WELCOME

SESSION 2

SPONSORED BY



FAIR DATA STEWARDSHIP AWARENESS COURSE

FAIR DATASTEWARDSHIP: AN INTRODUCTION





INTRODUCTION TO FAIR DATA STEWARDSHIP

FIRST THE FAIR DATA STEWARDSHIP PART

UCLA 1992



ERIK'S PHD PROJECT

Downloaded from rnajournal.cshlp.org on July 10, 2016 - Published by Cold Spring Harbor Laboratory Press

795

Global similarities among disparate ssRNA

TABLE 1. Phylogenetically and functionally representative compilation of ssRNA nucleotide composition.

RNA	Taxon	nª	$\langle N \rangle^{\rm b}$	$\langle G+C \rangle^c$	$\langle G+A \rangle^c$	$\langle G\!+\!U angle^c$
Comprehensive	_	2,800	287.0 ± 662.0	0.509 ± 0.200	0.516 ± 0.0310	0.523 ± 0.0462
235 rRNA	Archaea	15	$2,968.9 \pm 65.8$	0.588 ± 0.0525	0.567 ± 0.00597	0.509 ± 0.0212
200 110 111	Bacteria	39	$2,915.6 \pm 86.3$	0.526 ± 0.0383	0.570 ± 0.00720	0.517 ± 0.0102
	Eucarya	33	$3,615.0 \pm 470.3$	0.530 ± 0.0832	0.540 ± 0.0139	0.520 ± 0.0143
18S rRNA	Metazoa	20	$1,821.0 \pm 43.2$	0.494 ± 0.0297	0.517 ± 0.0123	0.535 ± 0.0113
16S rRNA	Archaea	15	$1,530.7 \pm 184.9$	0.611 ± 0.0427	0.562 ± 0.00500	0.507 ± 0.00681
100 110 110	Bacteria	85	$1.511.8 \pm 30.9$	0.550 ± 0.0387	0.568 ± 0.00770	0.520 ± 0.0118
	Eucarya	47	$1,823.6 \pm 57.8$	0.486 ± 0.0255	0.524 ± 0.00671	0.527 ± 0.00955
55 PRNA	Archaea	26	124.5 ± 3.9	0.598 ± 0.0726	0.508 ± 0.0272	0.498 ± 0.0409
55 IRIVA	Bacteria	123	117.6 ± 3.9	0.575 ± 0.0574	0.520 ± 0.0308	0.495 ± 0.0344
	Eucarva	234	119.3 ± 1.4	0.557 ± 0.0369	0.517 ± 0.0119	0.505 ± 0.0176
D DNIA	Archaga	7	400.9 ± 62.0	0.644 ± 0.0966	0.564 ± 0.0140	0.469 ± 0.0296
PRNA	Restaria	27	400.9 ± 02.0 280.0 ± 20.7	0.549 ± 0.0000	0.504 ± 0.0140 0.570 ± 0.0168	0.496 ± 0.0147
	Dacteria	37	569.0 ± 59.7	0.009 = 0.114	0.570 = 0.0100	0.100 - 0.0010
Group I ribozymes	_	13	757.15 ± 517.5	0.434 ± 0.105	0.561 ± 0.0302	0.492 ± 0.0268
Group II ribozymes	- 2000	6	734.8 ± 853.8	0.355 ± 0.0933	0.566 ± 0.0259	0.495 ± 0.0276
HDV Ribozymes	-	2	43.5 ± 43.1	0.619 ± 0.00212	0.512 ± 0.0368	0.519 ± 0.0269
snRNA, U1	Eucarya	24	162.4 ± 4.2	0.557 ± 0.0228	0.487 ± 0.00957	0.548 ± 0.0152
snRNA, U2	Eucarya	16	187.8 ± 11.6	0.455 ± 0.0357	0.456 ± 0.0310	0.543 ± 0.0257
snRNA, U3	Eucarya	8	220.6 ± 14.6	0.468 ± 0.0684	0.477 ± 0.0417	0.561 ± 0.0207
snRNA, U4	Eucarya	11	143.6 ± 12.3	0.484 ± 0.0384	0.493 ± 0.0192	0.536 ± 0.0303
snRNA, U5	Eucarya	11	125.6 ± 29.7	0.411 ± 0.0345	0.448 ± 0.0279	0.531 ± 0.0220
snRNA, U6	Eucarya	14	101.5 ± 7.7	0.451 ± 0.0233	0.544 ± 0.0304	0.469 ± 0.0352
tRNA	Comprehensive	2,011	74.3 ± 6.3	0.496 ± 0.120	0.510 ± 0.0280	0.528 ± 0.0509
	Archaea	121	77.0 ± 4.2	0.633 ± 0.0457	0.503 ± 0.0215	0.516 ± 0.0351
	Bacteria	371	78.3 ± 5.2	0.580 ± 0.0522	0.506 ± 0.0222	0.522 ± 0.0317
	Eucarva	436	75.8 ± 4.2	0.572 ± 0.0438	0.510 ± 0.0282	0.544 ± 0.0365
	Chloroplast	291	75.6 ± 5.0	0.526 ± 0.0531	0.510 ± 0.0212	0.540 ± 0.0341
	Mitochondria	742	70.3 ± 6.4	0.372 ± 0.0939	0.514 ± 0.0328	0.519 ± 0.0685
Artificial ribozymes ^d	Class I					
,	b1	1	274	0.471	0.511	0.493
	b1-207	1	119	0.513	0.546	0.445
	Class II					
	c2	1	271	0.568	0.535	0.546
	d1	1 .	273	0.487	0.502	0.524
	f1	1	273	0.546	0.524	0.502
	Class III					
	e3	1	272	0.529	0.511	0.563
	g1	1	274	0.544	0.522	0.518
Random		500	25	0.498 ± 0.104	0.495 ± 0.0979	0.494 ± 0.130
	_	500	74	0.500 ± 0.0560	0.502 ± 0.0587	0.498 ± 0.0548
	_	500	120	0.500 ± 0.0471	0.501 ± 0.0457	0.500 ± 0.0446
	_	500	400	0.501 ± 0.0260	0.501 ± 0.0247	0.500 ± 0.0254
	- 0100000	500	1,500	0.500 ± 0.0134	0.500 ± 0.0132	0.500 ± 0.0129
	- 0.000	500	3,000	0.500 ± 0.00924	0.500 ± 0.00954	0.500 ± 0.00888
	_	500	5,000	0.500 ± 0.00692	0.500 ± 0.00758	0.500 ± 0.00703

Downloaded from rnajournal.cshlp.org on July 10, 2016 - Published by Cold Spring Harbor Laboratory Press







FIGURE 2. (Legend on facing page.)

ERIK'S PHD DATA STEWARDSHIP PLAN



ERIK'S PHD DATA STORAGE PLAN



ERIK'S PHD DATA STORAGE PLAN SPECIFIED LOCATION



EUREKA





minus the background levels observed in the HSP in the control (Sar1-GDP-containing) incubation that prevents COPII vesicle formation. In the microsome control, the level of p115-SNARE associations was less than 0.1%.

- 46. C. M. Carr, E. Grote, M. Munson, F. M. Hughson, P. J. Novick, J. Cell Biol. 146, 333 (1999). 47. C. Ungermann, B. J. Nichols, H. R. Pelham, W. Wick-
- ner, J. Cell Biol. 140, 61 (1998). 48. E. Grote and P. J. Novick, Mol. Biol. Cell 10, 4149
- (1999). 49. P. Uetz et al., Nature 403, 623 (2000).
- 50. GST-SNARE proteins were expressed in bacteria and purified on glutathione-Sepharose beads using stan dard methods. Immobilized GST-SNARE protein (0.5 μM) was incubated with rat liver cytosol (20 mg) or purified recombinant p115 (0.5 µM) in 1 ml of NS buffer containing 1% BSA for 2 hours at 4°C with rotation. Beads were briefly spun (3000 rpm for 10 s) and sequentially washed three times with NS buffer and three times with NS buffer supplemented with 150 mM NaCl. Bound proteins were eluted three times in 50 µl of 50 mM tris-HCl (pH 8.5), 50 mM reduced glutathione, 150 mM NaCl, and 0.1% Triton

REPORTS

gel electrophoresis (PAGE) followed by immunoblot

52. K. G. Hardwick and H. R. Pelham, J. Cell Biol. 119, 513

53. A. P. Newman, M. E. Groesch, S. Ferro-Novick, EMBO

54. A. Spang and R. Schekman, J. Cell Biol. 143, 589 (1998).

55. M. F. Rexach, M. Latterich, R. W. Schekman, J. Cell

56. A. Mayer and W. Wickner, J. Cell Biol. 136, 307 (1997).

57. M. D. Turner, H. Plutner, W. E. Balch, J. Biol. Chem.

58. A. Price, D. Seals, W. Wickner, C. Ungermann, J. Cell

59. X. Cao and C. Barlowe, J. Cell Biol. 149, 55 (2000).

60. G. G. Tall, H. Hama, D. B. DeWald, B. F. Horazdovsky,

ting using p115 mAb 13F12.

(1992).

/. 11. 3609 (1992).

272, 13479 (1997).

Biol. 126, 1133 (1994).

Biol. 148, 1231 (2000).

Biol. Cell 8, 1089 (1997).

Mol. Biol. Cell 10 1873 (1999) 61. C. G. Burd, M. Peterson, C. R. Cowles, S. D. Emr. Mol.

51. V. Rybin et al., Nature 383, 266 (1996).

- X-100 for 15 min at 4°C with intermittent mixing, 62. M. R. Peterson, C. G. Burd, S. D. Emr, Curr. Biol. 9, 159 and elutes were pooled. Proteins were precipitated by (1999).
- MeOH/CH₃Cl and separated by SDS-polyacrylamide 63. M. G. Waters, D. O. Clary, J. E. Rothman, J. Cell Biol. 118, 1015 (1992).
 - 64. D. M. Walter, K. S. Paul, M. G. Waters, J. Biol. Chem. 273, 29565 (1998).
 - 65. N. Hui et al., Mol. Biol. Cell 8, 1777 (1997)
 - 66. T. E. Kreis, EMBO J. 5, 931 (1986).
 - 67. H. Plutner, H. W. Davidson, J. Saraste, W. E. Balch, . Cell Biol. 119, 1097 (1992).
 - 68. D. S. Nelson et al., J. Cell Biol. 143, 319 (1998).
 - 69. We thank G. Waters for p115 cDNA and p115 mAbs: G. Warren for p97 and p47 antibodies: R. Scheller for rbet1, membrin, and sec22 cDNAs; H. Plutner for excellent technical assistance; and P. Tan for help during the initial phase of this work. Supported by NIH grants GM 33301 and GM42336 and National Cancer Institute grant CA58689 (W.E.B.), a NIH National Research Service Award (B.D.M.), and a Wellcome Trust International Traveling Fellowship (B.B.A.).

20 March 2000; accepted 22 May 2000

One Sequence, Two Ribozymes: Implications for the Emergence of New Ribozyme Folds

Erik A. Schultes and David P. Bartel*

We describe a single RNA sequence that can assume either of two ribozyme folds and catalyze the two respective reactions. The two ribozyme folds share no evolutionary history and are completely different, with no base pairs (and probably no hydrogen bonds) in common. Minor variants of this sequence are highly active for one or the other reaction, and can be accessed from prototype ribozymes through a series of neutral mutations. Thus, in the course of evolution, new RNA folds could arise from preexisting folds, without the need to carry inactive intermediate sequences. This raises the possibility that biological RNAs having no structural or functional similarity might share a common ancestry. Furthermore, functional and structural divergence might, in some cases, precede rather than follow gene duplication.

Related protein or RNA sequences with the same folded conformation can often perform very different biochemical functions, indicating that new biochemical functions can arise from preexisting folds. But what evolutionary mechanisms give rise to sequences with new macromolecular folds? When considering the origin of new folds, it is useful to picture, among all sequence possibilities, the distribution of sequences with a particular fold and function. This distribution can range very far in sequence space (1). For example, only seven nucleotides are strictly conserved among the group I selfsplicing introns, yet secondary (and presumably tertiary) structure within the core of the ribozyme is preserved (2). Because these dispar-

*To whom correspondence should be addressed. Email: dbartel@wi.mit.edu

ate isolates have the same fold and function, it is thought that they descended from a common ancestor through a series of mutational variants that were each functional. Hence, sequence heterogeneity among divergent isolates implies the existence of paths through sequence space that have allowed neutral drift from the ancestral sequence to each isolate. The set of all possible neutral paths composes a "neutral network," connecting in sequence space those widely dispersed sequences sharing a particular fold and activity, such that any sequence on the network can potentially access very distant sequences by neutral mutations (3-5).

Theoretical analyses using algorithms for predicting RNA secondary structure have suggested that different neutral networks are interwoven and can approach each other very closely (3, 5-8). Of particular interest is whether ribozyme neutral networks approach each other so closely that they intersect. If so, a single sequence would be capable of folding into two different conformations, would

have two different catalytic activities, and could access by neutral drift every sequence on both networks. With intersecting networks, RNAs with novel structures and activities could arise from previously existing ribozymes, without the need to carry nonfunctional sequences as evolutionary intermediates. Here, we explore the proximity of neutral networks experimentally, at the level of RNA function. We describe a close apposition of the neutral networks for the hepatitis delta virus (HDV) self-cleaving ribozyme and the class III self-ligating ribozyme.

In choosing the two ribozymes for this investigation, an important criterion was that they share no evolutionary history that might confound the evolutionary interpretations of our results. Choosing at least one artificial ribozyme ensured independent evolutionary histories. The class III ligase is a synthetic ribozyme isolated previously from a pool of random RNA sequences (9). It joins an oligonucleotide substrate to its 5' terminus. The prototype ligase sequence (Fig. 1A) is a shortened version of the most active class III variant isolated after 10 cycles of in vitro selection and evolution. This minimal construct retains the activity of the full-length isolate (10). The HDV ribozyme carries out the site-specific self-cleavage reactions needed during the life cycle of HDV, a satellite virus of hepatitis B with a circular, single-stranded RNA genome (11). The prototype HDV construct for our study (Fig. 1B) is a shortened version of the antigenomic HDV ribozyme (12), which undergoes self-cleavage at a rate similar to that reported for other antigenomic constructs (13, 14).

The prototype class III and HDV ribozymes have no more than the 25% sequence identity expected by chance and no fortuitous structural similarities that might favor an intersection of their two neutral networks. Nevertheless, sequences can be designed that simultaneously satisfy the base-pairing requirements

Whitehead Institute for Biomedical Research and Department of Biology, Massachusetts Institute of Technology, 9 Cambridge Center, Cambridge, MA 02142, 1154

Α

	-	14.00	-			-	-				10/1	-		-	10/5					-	
	P1	J1/2	P2_	J2/		P1	P3	L3		P3	J3/4	P4		_P2	J2/5	P5	L5	_P5	J5/4	_ P4	
1		10		2	° —		30		4	2		50		60		-	70	8			
i.,		6 C 1	C P C	a	ale a	CHC.		00 10 11 11	u o olo	aluda	00.00	110001	1.2.0.2	Round	a altr	N O OFIC		Lu a calado	0.0	111 1 0 0 0	duo e
222	CCAGUC	CCAR	C al cl		den	CUG	a a a a a	CCUUU	nado		CCAC	UCCCI	1202	1000	C C U -	A 0 0 0 0 0	0000-	U A GALO	60-0	UAGGC	CLIGAR
1222	CCAGUC	GGAZ	C al cli		dea	CUG	ac a dd	CCUUU	TICCC	clu do	GGAG	UGCCI	1 A G A	ACUG	C C U -	ACCUC		UACACO		UNCCC	CLIGADE
1 2 2 2	CCAGUC	GGAN	C A CI		dea	CUG	C A C C	CCUUU	TIC CC	clu do	GGAG	UGCCI	I A G A	alcud	C C U -	adduc		UACACC	1	UAGGC	CLIGADA
AAA	CCAGUC	GGAZ	CACI	TATT	LAGA	CUG	ac a c c	CCUUU	TIGGG	au aa	GGAG	UGCCI	TAGA	AGUG	G G U -	aganc		UAGACO	A A - C	UAGGC	C LIG38
AAA	CCAGUC	GGAZ	CACE	CAUL	JAGA	CUGO	GCACC	CCUUUU	udda	gu d g	GGAG	UGCCI	LAGA	GGUG	G G U -	GGGUC	UUUU-	UAGACO	A A - C	UAGGC	C LIG36
AAA	CCAGUC	GGAZ	CACO	CAUL	JAGA	CUGO	GCACC	CCUUU	UGGG	GUGG	GGAG	UGCCU	JAGA	GGUG	GGU -	GGGUC	UUUUC	UAGACO	AA-C	UAGGC	LIG34
AAA	CCAGUC	GGAA	CACO	CAUL	JAGA	CUGO	GCACC	CCUUU	UGGG	GUGG	GGAG	uuccu	JAGA	GGUG	GGU-	GGGUC	UUUUC	UAGACO	AA-C	UAGGA	LIG32
AAA	CCAGUC	GGAA	CACO	CAUL	JAGA	CUGO	GCACC	CCUUU	UGGG	GUGG	GGAG	UUCCU	JAGA	GGUG	GGU-	GAGUC	UUUUC	UAGACI	AA-C	UAGGA	LIG30
AAA	CCAGUC	GGAA	CACO	CAUL	JAGA	CUGO	GCACC	CCUCC	UGGG	GUGG	GGAG	UUCCU	JAGA	GGUG	GGU-	GAGLUC	UUUUC	UAGACU	AA-C	UAGGA	LIG28
AAA	CCAGUC	GGAA	CACO	CAUL	JAGA	CUGO	GCACC	CCUCC	UGGG	GUGG	GGAG	UUCCU	JAGA	GGUG	GGU-	GAGCC	UUUUC	UAGGCU	AA-C	UAGGA	LIG26
AAA	CCAGUC	GGAA	CACO	CAUL	JAGA	CUGO	GCACG	CCUCC	UGGC	GUGG	GGAG	UULCC	JAGA	GGUG	GGU-	GAGCC	UUUUC	UAGGCU	AA-C	UAGGA	LIG24
AAA	CCAGUC	GGAA	CACO	CAUL	JAGA	CUGO	GCACG	CCUCC	UGGC	GUGG	GGAG	UUGLCU	JAGA	GGUG	G G U -	GAGCC	UUUUC	UAGGCU	AA-C	UAGCA	LIG22
AAA	CCAGUC	GGAA	CACO	CAUL	JAGA	CUGO	GCACG	CCUCC	UGGC	GUGG	GGAG	UUGGI	LAGA	GGUG	G G U -	GAGCC	UUUUC	UAGGCU	AA-C	UACCA	LIG20
AAA	CCAGUC	GGAA	CACO	CAUL	JAGA	CUGO	GCACG	CCUCC	UGGC	GUGG	GGAG	UUGGU	CGA	GGUG	GGU-	GAGCC	UUUUC	UAGGCU	AA-C	GACCA	LIG18
AAA	CCAGUC	GGAA	CACO	CAUL	JAGA	CUGO	GGACG	CCUCC	UGGC	GUCG	GGAG	UUGGU	CGA	GGUG	GGU-	GAGCC	UUUUC	UAGGCU	AA-C	GACCA	LIG16
AAA	CCAGUC	GGAA	CACO	CAUL	JAGA	CUGO	GGACG	CCUCC	UGGC	GLUCG	GGAG	UUGGO	GCGA	GGUG	G G U -	GAGCC	UUUUC	UAGGCU	AA-C	GCCCA	LIG14
AAA	CCAGUC	GGAA	GACO	CAUL	JAGA	CUGO	GGCCG	CCUCC	UGGC	GGCG	GGAG	UUGGO	GCGA	GGUG	GGU-	GAGCC	UUUUC	UAGGCU	AA-C	GCCCA	LIG12
AAA	CCAGUC	GGAZ	ULACO	CAUL	JAGA	CUGO	GGCCG	CCUCC	UGGC	GGCG	GGAG	UUGGO	GCGA	GGLUA	G G U -	GAGCC	UUUUC	UAGGCU	AA-C	GCCCA	LIG10
AAA	CCAGUC	GGAA	UCCO	CAUL	JAGA	CUGO	GGCCG	CCUCC	UGGC	GGCG	GGAG	UUGGO	GCGA	GGGA	G GLU -	GAGCC	UUUUC	UAGGCU	AA-C	GCCCA	LIG8
AAA	CCAGUC	GGAA	UCCO	CAUL	JAGA	CUGO	GGCCG	CCUCC	UGGC	GGCG	GGAG	UUGGO	CGA	GGGA	GGAA	GAGCC	UUUUC	UAGGCU	AA-C	GCCCA	LIG6
LA A A	CCAGUC	GGAA	UCCO	CAUL	JAGA	CUGO	GGCCG	CCUCC	UGGC	GGCG	GGAG	UUGGO	GCGA	GGGA	GGAA	CAGCC	UUUUC	UAGGCU	AA-C	GCCCA	LIGS
GAA	CCAGUC	GGAA	UCCO	CAUL	JAGA	CUGO	GGCCG	CCUCC	UGGC	GGCG	GGAG	UUGGO	3 CLG A	GGGA	GGAA	CAGCC	UUUUC	UAGGCU	AA-LC	GCCCA	LIG4
GAA	CCAGUC	GGAA	UCCO	CAUL	JAGA	CUGO	GGCCG	CCUCC	UGGC	GGCG	GGAG	UUGGO	GCUA	GGGA	GGAA	CAGCC	UUUUC	UAGGCU	A A - G	GCCCA	LIG2
GAA	CCAGUC	GGAL	UCCO	CAUL	JAGA	CUGO	GGCCG	CCUCC	UCGC	GGCG	GGAG	UUGGO	JCUA	GGGA	GGAA	CAGCC	UUUUC	UAGGCU	A A - G	GCCCA	LIG1
GAA	CCAGUC	GGAC	CCCC	CAUL	JAGA	CUGO	GGCCG	CCUCC	UCGC	GGCG	GGAG	UUGGO	GCUA	GGGA	GGAA	CAGCC	uuuuc	UAGGCU	A A - G	GCCCA	INT
GAA	CCAGUC	GGAC	CUCCO	CAUL	JAGA	CUGO	GGCCG	CCUCC	UCGC	GGCG	GGAG	UUGGG	3 C U A	GGGA	GGAA	CAGCC	UUUCC	UAGGCU	A A - G	GCCCA	HDV1
GAA	CCAGUC	- GAC	: UCCC	CAUL	JAGA	CUGO	GGCCG	CCUCC	UCGC	GGCG	GGAG	UUGGO	GCUA	GGGA	GGAA	CAGCC	UUUCC	UAGGCU	A A - G	GCCCA	HDV2
GGA	CCAUUC	- GAC	CCCC	CAUL	JAGA	CUGO	GGCCG	CCUCC	UCGC	GGCG	GGAG	UUGGO	3 C U A	GGGA	GGAA	CAGCC	UUUCC	UAGGCU	AA-G	GGCCA	HDV4
GGA	CCAUUC	- GAC	UCCO	CAUL	JAGA	CUGO	GUCCG	CCUCC	UCGC	GGCG	GGAG	UUGGO	JCUA	GGGA	GGAA	CAGCC	0 0 0 0 C C	UAGGCU	AA-G	GACCA	HDV6
GGA	CCAUUC	GAC	C U C C C	SAUL	JAGA	CUGO	GUCCG	CCUCC	UCGC	GGCG	GGAG	UUGGG	JCUA	GGGA	GGAA	CAGCC	U UIC CIC	UAGGCU	AA-G	GACCA	HDV7
GGA	CCAUUC	GAC	UCCIC	GAUL	JAGA	CUGO	GUCCG	CCUCC	UCGC	GGCC	GGAG	UUGGO	JCUA	GGGA	GGAA	CAGCC	UUCCC	UAGGCU	AA-G	GACCA	HDV9
GGA	CCAUUC	GAC	UCGG	GAUL	JAGA	CUGG	GUCCG	CCUCC	UCGC	GGCC	GGAG	UUGGG	CUA	GGGA	GGAA	CAGCC	U U C C C	UAGGCU	AA-G	GACCA	HDV11
GGA	CCAUUC	GAC	UCIGO	JAUL	JAGA	CUGC	GUCCG	CCUCC	UCGC	GGCC	GGAG	UUGGO	CALA	GGGA	GGAA	CAGCC	U UIC CIC	UJUGGCU	AA-G	GACCA	HDV13
GGA	CCAUUC	1000	UCGG	AUL	AGA	CUGG	GUCCG	CCUCC	UCGC	GGCC	CGAG	ducad	SCAU	GGGA	GGAA	CAGCC		AUGGCU	AA-G	GACCA	HDV15
COA	CCAUUC	1222	U C C C		JAGA	CUGG	00000	CCUCC	UCGC	00000	COAG	queec	CAU	CCCA	aca:	CAGCC		AUGGCU	A A - G	GACCA	HDV17
ada	CCAUUC	1000	U C C C		IAGA	CUGG	an c c a	CCUCC	UCGC	00000	C C A C	queed	CAU	CCCA	A G G A	CAGCC		AUGGCC	A A - 0	CACCA	HDW21
dda	CCAUUC		u clo	C AL	IACA	CUG	and coo	CCUCC	UCGC	cede	COAC	duced	CAU	CCCA	ACCA	CAGCC		AUGGCI	A A - 0	CACCA	HDV21
ada	CCAUUC		uca	SCAL	í 🗋 d	CUG	aucca	CCUCC	UCGC	aadd	CGAC	ducco	CAU	GGGA	AGGA	CAGCC		AUGGCU	A A - 0	GACCA	HDV25
ada	CCAUUC		ucla	CAL		C U G	CHCCG	CCUCC	UCGC	aada	CGAC	duce	CAU	GGGA	AGGA	CAGCC	u u c c c	AUGGCU	A A - 0	GAGCA	HDV27
GGA	CCAUUC	- G G G	ucla	CAL	il - G G	CUG	CHCCG	CCUCC	UCGC	aado	CGAC	duag	CAU	GGGA	AGGII	DAGCC	uuc cc	AUGGCU	AA - G	GAGCA	HDV29
GOA	CCAUUC		ucla	CAL	I - G G	CUG	cucca	CCUCC	UCGC	GGCC	CGAC	dugg	CAU	GGGA	AGGU	UAGCC	uuc de	AUGGCU	AAGG	GAGCA	HDV30
GGA	CCAUUC	- G G G	UCGO	CAL	I - GG	CUG	cuc ca	CCUCC	UCGC	G GIU C	CGAC	dugg	CAU.	GGGA	AGGU	HAGCC	uuccc	AUGGCU	AAGG	GAGCA	HDV32
GGA	CAUUC		UCGO	GCAL	I - G G	CUG	CUCCA	CCUCC	UCGC	GGUC	CGAC	duggo	CAU	GGGA	AGGU	UAGCC	uuccc	AUGGCU	AAGG	GAGCA	HDV33
GGA	C-AUUC	- G G G	UCGO	GCAL	I - GG	CUG	CUCCA	CCUCC	UCGC	GGUC	CGAC	dugg	CAU	GCIGA	AGGU	UAGCC	UUCGC	AUGGCU	AAGG	GAGCA	HDV34
GGA	C-AUUC	- G G G	UCGO	GCAL	J - G G	CUG	CUCCA	CCUCC	UCGC	GGUC	CGAC	dugg	CAU	GCGA	AGGU	UUUCC	UUCGC	AUGGCU	AAGG	GAGCA	HDV36
GGA	C-AUUC	- G G G	UCGO	GCAL	J - G G	CUG	CUCCA	CCUCC	UCGC	GGUC	GAC	duggg	CAU	CCGA	AGGU	UUUCC	UUCGG	AUGGCU	AAGG	GAGCA	HDV38
GGA	C-AUUC	- G G G	UCGO	GCAL	J - G G	CUU	CUCCA	CCUCC	UCGC	GGUC	CGAC	duggo	CAU	CCGA	AGGU	UUUCC	UUCGG	AUGGCU	AAGG	GAGAA	HDV40
GGA	C-AUUC	- G G G	UCGO	GCAL	J - G G	CAUC	CUCCA	CCUCC	UCGC	GGUC	GGAC	CUGGO	CAU	CCGA	AGGU	UUUCC	UUCGG	AUGGCU	AAGG	GAGAG	HDV42
GGG	A-AUUC	- GGG	UCGO	SCAL	1-00	CAUC	CUCCA	CCUCC	UCGC	GGUC	CGAC	duggo	CAU	CCGA	AGGU	UUUCC	UUCGG	AUGGCU	AAGG	GAGAG	HDV P
		and so the second				and the second second				-	-		Bolevalue	-			-				
		P	1		1/2	P2	2 F	23	L3	P3	P1			P4		L4	P4		1/2	P2	

AND IMAGES



ERIK'S POSTDOC DATA STEWARDSHIP PLAN



- 24. M. A. Tanner and T. R. Cech, RNA 2, 74 (1996).
- 25. Supplemental data showing the predicted secondary structures of each construct (Fig. 3) and explaining the ligation activity of truncated ribozymes (Fig. 2B) are available at Science Online at www.sciencemag. org/feature/data/1050240.shl.
- 26. J. Maynard Smith, Nature 225, 563 (1970).
- 27. The Intersection Theorem (5) states that there exists at least one intersection sequence for every pair of RNA secondary structures (usually as suboptimal conformations).
- S. Ohno, Evolution by Gene Duplication (Springer-Verlag, New York, 1970).
- 29. D. L. Minor Jr. and P. S. Kim, Nature 380, 730 (1996).
- M. H. J. Cordes, N. P. Walsh, C. J. McKnight, R. T. Sauer, Science 284, 325 (1999).
- 31. X. Ye et al., Chem. Biol. 6, 657 (1999).

MMM: A 404!



BUT THERE IS GOOGLE NOW



https://www.utexas.edu/sites/default/files/Lambowitz%20Paper.pdf ▼ by H Guo - 2000 - Cited by 194 - Related articles Mar 12, 2013 - the ligation activity of truncated ribozymes (Fig. 2B) are available at Science Online at www.sciencemag.org/feature/data/1050240 sbl 26

SCIENCE MUST HAVE THEIR ACT TOGETHER?

From: Erik Schultes erik.schultes@dtls.nl Subject: Supplemental Data Date: October 1, 2016 at 11:20 AM To: science_editors@aaas.org

To whom it may concern -

I authored this article in 2000: http://science.sciencemag.org/content/289/5478/448.article-info

At the time, we had submitted Supplemental Data.

However, on the webpage supporting the paper, I no longer see a link to the Supplemental Data.

In endnote 25 of the paper, I see this link to the Supplemental Data: www.sciencemag.org/feature/data/l050240.shl

However, this link appears to be broken, and redirects to a 404 page: http://www.sciencemag.org/feature/data/i050240.shl

So I have a few guestions:

(1) Is it possible, one way or another, to obtain or access the original Supplemental Data ?

(2) My immediate interest is not so much at the content level, but more at the 'Data Stewardship' level. Recently, my colleagues and I conducted a LERU Summer School on Data Stewardship (in Leiden, the Netherlands, see http://www.dtls.nl/fair-data/fair-data/fair-data/fair-data/fair-data/fair-data/fair-data/fair-data/fair-data/fair-data/fair-data/fair-data-training/leru-summer-school/). I actually used this 'broken link' as a case study for the kinds of problems we all face in scholarly data stewardship. So I'm wondering, can you tell me when and why the link was broken ? It is instructive for us to better understand the challenges that publishers face these days.

Many thanks in advance,

Best regards,

Erik

Erik Schultes PhD FAIR Data Scientific Projects Lead

Dutch Techcentre for Life Sciences Visiting address: Catharijnesingel 54 | 3511 GC Utrecht Postal address: Postbus 19245 | 3501 DE Utrecht

E-mail: erik.schultes@dtls.nl Skype: easchultes Tel: +31 642448027 Website: www.dtls.nl

NO REPLY TO DATE!



Instance One Sequence, Two Ribozymes: Implications for the Emergence of New Ribozyme Folds FULL TEXT CErik A. Schultes and David P. Bartel

(827K JPEG file)]

To Advertise Find Products

Supplementary Material

Supplemental Figure 1. Secondary structures of the ribozyme variants aligned in Fig. 3A. Accumulated changes from the intersection sequence are indicated (red and blue residues, with blue identifying residues changed at the step indicated by an arrow; dashes mark sites of point deletions). Self-ligation and self-cleavage rates (min⁻¹) are listed below the name of each construct.

Larger, segmented version of this image (new window - 160K JPEG images)

Full version of this image Download PDF of this image (9MB)

Explanation of the ligation activity of truncated ribozymes.

For LIG2, LIG1, INT, and HDV1, the ligation products are represented by two bands in Fig. 2B (top panel), due to unanticipated ligation activity of truncated ribozyme molecules in addition to the activity of full-length RNA. On lower percentage gels, the faster-migrating product band resolves into seven distinct bands, corresponding to a loss of 9 to 15 nucleotides at the ribozyme 3' terminus. These truncated ribozymes appear to adopt a class III-like fold as evidenced by the following: (i) They produce 2'-linkages. (ii) Nucleotide substitutions known to favor the class III ligase function and fold (C13A and C38G) lead to parallel increases in the activity of the truncated RNAs (Fig 2). (iii) Constructs LIG2, LIG1, INT, and HDV1 have fortuitous potential base pairing between segments corresponding to G51-U55 and A66-U70 in Fig. 2A. For the truncated molecules, this base pairing would lead to a fold that has all the class III features except the P4 stem-loop (Web fig. 2). The fortuitous potential pairing is not present in the original class III ligase isolates or in the constructs of the ligase neutral path (LIG4-LIG P in Fig. 3), explaining why only one product band was observed for LIG4-LIG P. (iv) The Pb(II)-cleavage pattern of the ligation product of truncated LIG2 was consistent with the formation of G51-U55:A66-U70 pairing, whereas probing of full-length LIG2, LIG1, and INT products confirmed formation of the anticipated P4 and P5 stems. None of the fortuitous pairs are present in the HDV secondary structure, and thus the alternative ligase fold, like the prototype ligase fold, has no base pairs in common with the HDV fold. The 3' truncations contaminating the ribozyme preparations of LIG2, LIG1, INT, and HDV1 can generate over half of the ligated product (Fig. 2B), even though they compose only a minor fraction of the ribozyme preparation (presumably arising from RNA degradation or premature transcription termination). For these truncated molecules, a more optimal P5 might more than offset the loss of P4.

Supplemental Figure 2. The intersection sequence and a truncated form of this sequence assuming the class III ligase fold and the class III ligase-like fold, respectively.



Medium version | Full size version

Science. ISSN 0036-8075 (print), 1095-9203 (online)

xhtml org/site/feature/data/1050240. ciencemag. ш \cup 12:45 in October 10, http://www

FIRST: THE DATA STEWARDSHIP PART



Stewardship is an ethic that embodies the responsible planning and management of resources. Can be applied to nature, economics, health, property, information, theology, etc.

https://en.wikipedia.org/wiki/Stewardship

DMP: PROJECT

DSP: Long-term re-use

Data Stewardship: A plan for maximizing the re-use of data.

exposing versus publishing data

- 폐 hiding versus securing data
- 폐 open versus closed data
- licensing and fee
- authors versus owners (eg patient privacy)
- 🥶 who paid for the data ?
- 💿 big data scaling: storage, compute
- serving (e.g. 24/7) versus archiving
- machine interoperation (ontology engineering, data modeling)
- human interoperation (24 EU languages)

Data Stewardship: A plan for maximizing the re-use of data. Systematic organisational changes (policy, mandates, budgets)

🥯 New Technologies

Training for new profession

Data Stewardship: A plan for maximizing the re-use of data.



















End





FOSTER: practical implementation of Open Science in Horizon 2020 and beyond

https://www.fosteropenscience.eu/

EDISON Data Science Framework to define the Data Science Profession

http://edison-project.eu/edison

Home



Center of Open Science, providing tools, training, support and advocacy for changes in incentives to make research investment go farther, faster.

ine.dcc.ac.uk Sign in Email nline helps you to create, review, and share data management plans that and funder requirements. It is provided by the Digital Curation Centre (DCC). Join the growing international community that have adopted DMPonline Password Forgot password? 17.622 Users Remember email or Some funders mandate the use of DMPonline, while others point to it as a useful option. You can download funder templates without logging in, but the tool provides tailored guidance and example answers from the DCC and many research organisations. Why not sign up for an account and try it out? DCC © 2010 - 2018 Digital Curation Center Terms of use Privacy statement Github About Contact us Learn - Sign in - English (US) https://dmptool.org Sector DMP Tool Welcome Get started Create data management plans that meet institutional and funder requirements

🔯 Language -

Public DMPs Funder requirements Help

	and the second se		
DMPTool by the	Numbers		Top 5 Templates
88 30,169	26,667	234	Department of Energy (DOE): Generic Digital Curation Centre NSF-ENG: Engineering

https://cos.io

БМР

http

Some f

example

and try it

© 2010 - 20

Welcom

DMPTc

Create data ma



Center of Open Science, providing tools, training, support and advocacy for changes in incentives to make research investment go farther, faster.

https://cos.io

		Features	Business	Explore	Marketp	lace	Pricing		Sear	ch
5	DN	IPRoadma	p / roadma r)						
ne er gr	<> C	ode 🕛 I	ssues 79	ື່າ) Pull reque	sts 1	💷 Proj	ects 2	E Wil	<i< th=""><th>🔟 In</th></i<>	🔟 In
L	Ho	me								

stephaniesimms edited this page on Mar 27 · 27 revisions

roadmap

Welcome to the DMPRoadmap wiki!

The Digital Curation Centre and UC3 team at the California Digital Library have dever delivered tools for data management planning since the advent of open data policies (DCC-UK) and the DMPTool (CDL-US) are now established in our national contexts a resource for researchers seeking guidance in creating data management plans (DMF worked together from the outset to share experiences, but with the explosion of inter of our tools across the globe we formalized our partnership to co-develop and maint open-source platform for DMPs. By working together we can extend our reach, keep down, and move best practices forward, allowing us to participate in a truly global op ecosystem.

The new platform is separate from the services each of our teams runs on top of it. If enhancements will focus on making DMPs machine-actionable so please continue so use cases!

- Track our progress via the DMPonline blog and/or DMPTool blog
- Join the user community listserv: https://www.jiscmail.ac.uk/DMPONLINE-USEF
- Request an invitation to join the developer #community Slack team by emailing DMPonline@dcc.ac.uk with a note of the email addresses that should be added

DATA STEWARDSHIP: PLANNING TOOLS



The Data Stewardship Mind Map



DUTCH TOCHCENTRE FOR LIFE SCIENCES



THE DATA STEWARDSHIP MIND MAP



Rob Hooft

GETTING SYSTEMATIC ABOUT FAIR DATA STEWARDSHIP



Data Cycle Step 1: Design of Experiment Data Cycle Step 2: Data Design and Planning Data Cycle Step 3: Data Capture (Equipment) Data Cycle Step 4: Data Processing and Curation Data Cycle Step 5: Data Linking and Integration Data Cycle Step 6: Data Analysis and Interpretation Data Cycle Step 7: Publishing

DATA STEWARDSHIP: PLANNING TOOLS

CHAPTER 2

Data Cycle Step 1: Design of Experiment

Before you decide to embark on any new study, it is nowadays good practice to consider all options to keep the data-generation part of your study as limited as possible. It is not because we can generate massive amounts of data that we always need to do so. Creating data with public money brings with it the responsibility to treat those data well, and (if potentially useful) make them available for reuse by others. There is considerable effort and cost associated with making data FAIR, and generally speaking, recreating data that may exist somewhere else is a waste of public resources. So, given the research question you would like to address, the very first question in open science setting should always be:

1 IS THERE PRE-EXISTING DATA?

What's up?

For many decades if not centuries, virtually every experiment started with the ordection or creation of observations, and, in fact, data. In social sciences and humanities, the tendency to reuse data that had been created earlier, in all kinds of surveys and increasingly, of course, from sources such social media, may be already somewhat more established. However, in many of the hard experimental sciences, the generation of new data specifically produced to answer a hypothetical question is still so commonplace that careful thinking about the actual need to generate new data may just not be on the radar screen. Obviously,

64 Data Stewardship for Open Science: Implementing FAIR Principles

Re

2.2

What

Even i

matica

Section

or get

Howev

need to

• (

• If

DO

data creation will need to continue, but increasingly we have to ask the data creation will need to continue, a boolutely necessary to answer the question whether such new data are absolutely necessary to answer the question whether such new data are and more data becoming avail question we want to answer. With more and more data becoming avail able in reusable format, there may well be existing data collections of able in reusable format, increased services (OPEDAS) that with or other people's data and account with or without some extra effort, can answer at least part of the question or at least h ay be crucial for the interpretation of your own data

DO

- Search for datasets (OPEDAS) that may be reusable and can help ou reduce the number of new datasets you may have to scherate (and steward later on).
- Include annotated collections of data and curated databases in your search.
- Check the accessibility and license situation attached to the reevant datasets you found.
- Check their interoperability. They may be relevant but not interoperable with your analysis pipelines. In that case, you may have to extract, transform, and load (ETL) them, or decide that - although relevant - they are not reusable for your purpose.
- Ensure that using OPEDAS will not restrict in any way the used your results later on, including copyright and freedom to operate on the request of IPR.
- Check how to cite and acknowledge OPEDAS.
- Consider how to actively involve OPEDAS owners in your search, in order to make optimal use of their data.

k to colleagues who did similar experiments before, to fit out about potential OPEDAS you may consider using-DON'T

- Assume no OPEDAS exist without thoroughly checking and in a your possibilities • Start an experiment without properly checking with collest about the best approach.

about the best approach and OPEDAS out there.

GETTING SYSTEMATIC ABOUT FAIR DATA STEWARDSHIP

https://dsw.fairdata.solutions Erik Schultes (Common ELIXIR Knowledge Model, 1.0.0) DSW Save Ø Design of experiment **Current Phase** 3 Before you decide to embark on any new study, it is nowadays good practice to consider all options to keep the Before Submitting the Proposal \$ data generation part of your study as limited as possible. It is not because we can generate massive amounts of data that we always need to do so. Creating data with public money is bringing with it the responsibility to treat those data well and (if potentially useful) make them available for re-use by others. Design of experiment ✓ Is there any pre-existing data? Data design and planning **~** Are there any data sets available in the world that are relevant to your planned research? Data Capture/Measurement \checkmark Section 2017 Desirable: *Before Submitting the DMP* Data processing and curation \checkmark Data Stewardship for Open Science: *atq* Data integration \checkmark No ○ Yes \equiv Data interpretation Information and insight \checkmark Will reference data be created? Will any of the data that you will be creating form a reference data set for future research (by others)? Summary Report Solution Desirable: *Before Submitting the DMP* Data Stewardship for Open Science: *rbz* O No ○ Yes \equiv \equiv

Will you be storing samples?

☑ Desirable: *Before Submitting the DMP*

0

0

0

F٩

GETTING SYSTEMATIC ABOUT FAIR DATA STEWARDSHIP



ELIXIR Data Stewardship

Knowledge Model https://github.com/DataStewardshipWizard/ds-km

DS Wizard + Metrics Hackathon, Leiden, July 2-4





ELIXIR Data Stewardship

Knowledge Model https://github.com/DataStewardshipWizard/ds-km

GETTING SYSTEMATIC ABOUT FAIR DATA STEWARDSHIP

Data Stewardship Wizard	common ELIXIR (Common ELIXIR Knowledge Model, 1.0.0)	Save
	Data design and planning Answered: 54/54	
So KM Packages		
DS Planner	Metric Measure	
	Findability 0.33	
	Accessibility 0.25	
	Interoperability 0.63	
	Reusability 0.86	
	Good DMP 0.40 Practice	
	Openness 0.00	


ELIXIR Data Stewardship Knowledge Model



ELIXIR Data Stewardship Knowledge Model



ELIXIR Data Stewardship Knowledge Model

Findable:

F1 (meta)data are assigned a globally unique and persistent identifier;

F2 data are described with rich metadata;

F3 metadata clearly and explicitly include the identifier of the data it describes;

F4 (meta)data are registered or indexed in a searchable resource;

Interoperable:

11 (meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation.

I2 (meta)data use vocabularies that follow FAIR principles;

13 (meta)data include qualified references to other (meta)data;

Accessible:

A1 (meta)data are retrievable by their identifier using a standardized communications protocol;

A1.1 the protocol is open, free, and universally implementable;

A1.2 the protocol allows for an authentication and authorization procedure, where necessary;

A2 metadata are accessible, even when the data are no longer available;

Reusable:

R1 meta(data) are richly described with a plurality of accurate and relevant attributes;

R1.1 (meta)data are released with a clear and accessible data usage license;

R1.2 (meta)data are associated with detailed provenance;

R1.3 (meta)data meet domain-relevant community standards;

Measuring FAIRness http://fairmetrics.org

Framework for authoring FAIR Metrics

FAIR Metrics

The FAIR Metrics Group took-on the challenge of designing a framework for evaluating "FAIRness".

Discoverability and reusability are not abstract concepts, but imply concrete behaviors and properties that must hold true for the fulfillment of the FAIR objectives. Given this, it must therefore be possible to precisely define a measurable set of properties and behaviors that assess FAIRness. Over the short 1 month lifespan of the FAIR Metrics Working Group, we have created a cogent framework for developing FAIR metrics manifested as a simple form with 8 questions that structures fruitful conversations about proposed metrics.

Our approach recognizes that the diversity in opinion must play a key role in crafting fair and effective FAIR guidelines. Communities must not only understand what is meant by FAIR, but must also be able to monitor the FAIRness of their digital resources, in a realistic, but quantitative manner. We recognize that what is considered FAIR in one community may be quite different from FAIRness in another community - different community norms and practices make this a certainty! As such, our approach focuses on the mechanism by which metrics can be created by community members themselves, rather than attempting to create a set of one-size-fits-all metrics to apply to every resource.

FAIR Metrics GitHub FAIR Metrics Paper Metrics Process Metrics Authoring Framework Metrics Form

About Us

Measuring FAIRness http://fairmetrics.org

Framework for authoring FAIR Metrics

www.nature.com/scientificdata

SCIENTIFIC DATA

OPEN Comment: A design framework and exemplar metrics for FAIRness

Mark D. Wilkinson¹, Susanna-Assunta Sansone², Erik Schultes³, Peter Doorn⁴, Luiz Olavo Bonino da Silva Santos^{5,6} & Michel Dumontier⁷

Received: 28 November 2017 Accepted: 9 May 2018

The FAIR Principles¹ (https://doi.org/10.25504/FAIRsharing.WWI10U) provide guidelines for the publication of digital resources such as datasets, code, workflows, and research objects, in a manner that makes them Findable, Accessible, Interoperable, and Reusable (FAIR). The Principles have rapidly been adopted by publishers, funders, and pap-disciplinary infrastructure programmes and societies. The

Measu	uring FAIRness	FM	Question	Dataverse	Dryad	Nano- pub	Zenodo	Yale ISPS	Figshare	Broad's SCP	SeaData Net's CDI	Wikidata	trics
http://	fairmetrics.org	IRI Exists	1	IRI	IRI	IRI	IRI	IRI	IRI	IRI	IRI	IRI	
1		F1A	2	IRI	IRI	IRI	IRI	IRI	IRI	IRI	IRI	IRI	
		F1B	3	IRI	IRI	IRI	NRP	none	IRI	IRI	IRI	IRI	ta
		F2A	4A	IRI	IRI	IRI	IRI	none	none	IRI	IRI	IRI	
		F2A	4B	IRI	none	IRI	IRI	"Multiple"	none	IRI	IRI	IRI	
		F3	5A	IRI	IRI	IRI	IRI	none	NRP	IRI	IRI	IRI	
		F3	5B	IRI	IRI	IRI	IRI	IRI	IRI	IRI	none	IRI	
		F4	6A	IRI	IRI	IRI	IRI	IRI	IRI	IRI	IRI	IRI	101
		F4	6B	IRI	IRI	IRI	IRI	IRI	IRI	IRI	IRI	IRI	
		A1.1	7A	IRI	IRI	IRI	IRI	IRI	IRI	IRI	IRI	IRI	
		A1.1	7B	true	true	true	true	true	true	true	true	true	
	OP	A1.1	7C	true	true	true	true	true	true	true	true	true	DI
		A1.2	8A	false	false	false	false	false	false	false	true	false	
		A1.2	8B	N/A	N/A	N/A	N/A	NRP	NRP	NRP	link	N/A	
		A2	9	IRI	IRI	none	IRI	none	IRI	none	IRI	NRP	
		11	10	IRI	IRI	IRI	IRI	none	none	NRP	IRI	IRI	
		12	11	IRI	IRI	IRI	none	none	none	IRI	IRI	IRI	
		13	12	NRP	IRI	IRI	none	none	none	NRP	NRP	IRI	4.
	Received: 28 Novemb	R1.1	13	IRI	IRI	IRI	IRI	IRI	IRI	NRP	IRI	IRI	that
	Accepted: 9 M	R1.2	14A	IRI	IRI	IRI	IRI	none	none		NRP	NRP	been





THE PREPRINT SERVER FOR BIOLOGY

HOME | ABOUT | SUBMIT | ALERTS / RSS | CHANNELS

Search

Q

Advanced Search

New Results

Evaluating FAIR-Compliance Through an Objective, Automated, Community-Governed Framework

Mark D Wilkinson, D Michel Dumontier,
Susanna-Assunta Sansone, D Luiz Olavo Bonino da Silva Santos,
Mario Prieto, D Julian Gautier, D Peter McQuilton,



14 CORE METRICS

Findable:

F1 (meta)data are assigned a globally unique and persistent identifier;

FM-F1B

FM-F3

F2 data are described with rich metadata;

F3 metadata clearly and explicitly include the identifier of the data it describes; FM-F4

F4 (meta)data are registered or indexed in a searchable resource;

Interoperable:

I1 (meta)data use a formal, accessible, shared, and broadly applicable language for knowledge FM-I1 representation.

12 (meta)data use vocabularies that follow FAIR principles,

I3 (meta)data include qualified references to other (meta)data; FM-I3

Sci. Data 3:160018 doi: 10.1038/sdata.2016.18 (2016) http://fairmetrics.org https://github.com/FAIRMetrics/Metrics/blob/master/ALL.pdf

Accessible:

A1 (meta)data are retrievable by their identifier using a standardized communications protocol;

A1.1 the protocol is open, free, and universally implementable;

A1.2 the protocol allows for an authentication and authorization procedure, where necessary;

A2 metadata are accessible, even when the Fata are no longer available;

Reusable:

R1 meta(data) are richly described with a plurality of accurate and relevant attributes;

R1.1 (meta)data are released with a clear and accessible data usage license; FM-R1.1

R1.2 (meta)data are associated with detailed provenance;

FM-R1.2

R1.3 (meta)data meet domain-relevant community standards; FM-R1.3

FIELD	DESCRIPTION
Metric Identifier	FM-F1B: https://purl.org/fair-metrics/FM_F1B
Metric Name	Identifier persistence
To which principle does it apply?	F1
What is being measured?	Whether there is a policy that describes what the provider will do in the event an identifier scheme becomes depre- cated.
Why should we measure it?	The change to an identifier scheme will have widespread implications for resource lookup, linking, and data sharing Providers of digital resources must ensure that they have a policy to manage changes in their identifier scheme, with a specific emphasis on maintaining/redirecting previously generated identifiers.
What must be provided?	A URL that resolves to a document containing the relevant policy.
How do we measure it?	Use an HTTP GET on URL provided.
What is a valid result?	Present (a 200,202,203 or 206 HTTP response after resolving all and any prior redirects. e.g. 301 -> 302 -> 200 OK. or Absent (any other HTTP code)
For which digital resource(s) is	All
this relevant?	
Comments	A first version of this metric would focus on just checking a URL that resolves to a document. We can't verify that

FAIR Metrics Upgrades Example: FM-F1B, Identifier Persistence

v1.0 checks for HTTP 200 return

v2.0 validates a standard RDF persistence policy

v3.0 scores multiple parameters of persistence policy

The "15th" FAIR Metric



FAIR data and a new approach for data management September 21 2018 Den Haag

Nicoline Smit, Project Manager at Netherland Heart Institute Mira van der Naald, Department of Cardiology, UMC Utrecht

https://preclinicaltrials.eu



Preclinicaltrials aims to provide a comprehensive listing of preclinical animal study protocols.

Preferably registered at inception in order to **increase transparency**, help avoid duplication, and reduce the risk of reporting bias by enabling comparison of the completed study with what was planned in the protocol.

you to create an account that is

- Anonymous
- Free of charge
- Has an optional embargo period

This register is web-based, open to all types of animal studies and freely accessible and searchable to all with a prodinicaltrials ou account

experts on preclinical animal studies and preclinical evidence synthesis.

Please join us and create an user account, this will provide access to the database and enables you to register your preclinical trial.

Contact us

at info@preclinicaltrials.eu.



Section 1. General information

1. * Title of the study Enter the full title of the study

2. Acronym/short title Enter optional acronym/short title for the study

Preclinicaltrials comprehensive li animal study pro

Û

Ho

Preferably registe order to increase avoid duplicatic of reporting biase comparison of the what was planne 3. * Contact details Give the name of the main administrative contact for the study

Name

Role

What is the role the main contact in the study (e.g. executive researcher, research group supervisor)?

esigned by al studies nthesis.

witter

E⇒

an user ccess to the to register

Metric Identifier	FM-CT1 (FAIR Metric Clinical Trail 1)						
Metric Name	Preregistration						
To which principle does it apply?	R1.2 (meta)data are associated with detailed provenance						
What is being measured?	The existence of clinical trail preregistration						
Why should we measure it?	Preregistration is important for Increased transparency and reduced risk of bias and help avoid duplication.						
What must be provided?	A URL to the preclinical registration document						
How do we measure it?	Use HTTP GET on URL provided.						
What is a valid result?	HTTP 200 (now); Validted RDF file (later)						
For which digital resource(s) is this relevant?	preclinicaltrails.eu						

FAIR Metric

The "15th"

EEDAR retadatacentec.org	Search			۹			٠	٠
New +	All / Users / Marcos Ma	rtinez Romero / Preclinical Tr	rials form				i	11
Workspace	Title			Created	Modifi			
Shared with Me	PRECLINICALTR	RIALS.EU		2:33 PM	3:52 PM		•	1
	🗞 Blinding			3:29 PM	3:29 PM			1
FILTER RESET ALL		CALTRIALS.EU	.EU					
		1. Title of the study*	SECTION 1: G	ENERAL INFC	RMATI	NO		
		2. Acronym/short title						
		3. Contact Details Name*						
		Role*						
		4. Study centre details*						



ELIXIR Data Stewardship Knowledge Model



 \odot







Making it easy for humans to make metadata for machines

https://www.go-fair.org/resources/go-fair-workshop-series/



First M4M: October 15-16, Leiden Co-organizers: Wittenburg & Schultes

https://digitalscholarshipleiden.nl/articles/metadata-4-machines-help-you-find-and-reuse-relevant-research-data

Metadata for Machines

Vehicle for Community Decision Making

Likely Stakeholders:

- 🕯 Data steward
- Researcher / Research community
- University / organziations (NFU)
- **Generation** Funder
- Generation Publisher
- **General Repository**
- **Generation** FAIR tools and services
- Generation and other 3rd parties

GO FAIR Meta Data for Machines Pilot



















ELIXIR Data Stewardship Wizard

Research Communities

- preclinical research
- chemistry
- earth science







Write DS Plan







INTRODUCTION TO FAIR DATA STEWARDSHIP

SECOND: THE FAIR PART

"Data and services that are findable, accessible, interoperable, re-usable for machines and for people"

The FAIR Guiding Principles for scientific data management and stewardship, Scientific Data (2016), https://www.nature.com/articles/sdata201618

"Data and services that are findable, accessible, interoperable, re-usable for machines (and sometimes, in rare circumstances, may be even for people."

The FAIR Guiding Principles for scientific data management and stewardship, Scientific Data (2016), https://www.nature.com/articles/sdata201618

WHAT IS FAIR DATA: MORE DETAIL ON THE PRINCIPLES



GO FAIR Initiative Implementation Networks FAIR Principles Technology Training Certification

News Contact Q

FAIR Principles

Home > FAIR Principles

FAIR Principles

- F1: (meta) data are assigned globally unique and persistent identifiers
- F2: Data are described with rich metadata
- F3: Metadata clearly and explicitly include the identifier of the data it describes

On March 15th, 2016, a group has published "**The FAIR Guiding Principles for scientific data management and stewardship**" comment on Nature's Scientific Data. The authors aimed for the principles to act as guidelines for those willing to improve findability, accessibility, interoperability and reuse of their digital assets. With the increase on volume, complexity and creation speed of data, humans are more and more relying on computational support for dealing with data. The principles were, therefore, defined with the focus on machine-actionability, i.e., the capacity of computational systems to find, access, interoperate and reuse data with none or minimal human intervention.

Findable: the first obstacle for someone willing to (re)use data is to find them. Metadata

FAIR DATA PRINCIPLES

Findable:

F1. (meta)data are assigned a globally unique and persistent identifier;

F2. data are described with rich metadata;

F3. metadata clearly and explicitly include the identifier of the data it describes;

F4. (meta)data are registered or indexed in a searchable resource;

Interoperable:

11. (meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation.

I2. (meta)data use vocabularies that follow FAIR principles;

I3. (meta)data include qualified references to other (meta)data;

Accessible:

A1. (meta)data are retrievable by their identifier using a standardized communications protocol;

A1.1 the protocol is open, free, and universally implementable;

A1.2. the protocol allows for an authentication and authorization procedure, where necessary;

A2. metadata are accessible, even when the data are no longer available;

Reusable

R1. (meta)data are richly described with a plurality of accurate and relevant attributes;

R1.1. (meta)data are released with a clear and accessible data usage license;

R1.2. (meta)data are associated with detailed provenance;

R1.3. (meta)data meet domain-relevant community standards;

https://www.nature.com/articles/sdata201618

Sci. Data 3:160018 doi: 10.1038/sdata.2016.18 (2016)
OPEN SCIENCE TAXONOMY

Open Science Taxonomy



OPEN SCIENCE



Die 6 Prinzipien

Offene Wissenschaft basiert auf sechs Prinzipien, um Teilschritte und -ergebnisse eines wissenschaftlichen Prozesses zu öffnen. Die ersten vier Prinzipien basieren auf dem Paper "<u>The</u> <u>Case for an Open Science in Technology Enhanced Learning</u>" (<u>Kraker 2011</u>). Open Peer Review und Open Educational Resources sind zwei weitere wichtige Aspekte von Wissenschaft. Die "<u>Open Definition</u>" erklärt, was "offen" in diesem Kontext bedeutet.

- **Open Methodology:** das Anwenden von Methoden sowie den gesamten Prozess dahinter soweit praktikabel und relevant dokumentieren
- **Open Source:** Quelloffene Technologie (Soft- und Hardware) verwenden und eigene Technologien öffnen
- Open Data: Erstellte Daten frei zur Verfügung stellen
- **Open Access:** In einer offenen Art publizieren, und für jedeN nutzbar und zugänglich machen (<u>s. Budapest Initiative (eng)</u>)
- **Open Peer Review:** Transparente und nachvollziehbare Qualitätssicherung durch offenen Peer Review
- **Open Educational Resources:** Freie und offene Materialien für Bildung und in der universitären Lehre verwenden

A REFERENCE IMPLEMENTATION OF THE 15 PRINCIPLES

How FAIR Data can be created

 $((\cdot))$

(FAIR-dICT project, DTL: <u>https://www.dtls.nl/fair-data/fair-dict/</u>)



A REFERENCE IMPLEMENTATION OF THE 15 PRINCIPLES



FAIR DATA POINT



FAIR META DATA

FAIR metadata

litle	FDP of lorentz.fair-dtls.surf-hosted.nl
Metadata ID	3e77134d-9338-482a-a4f7-8e6933b47469
Description	FDP of lorentz.fair-dtls.surf-hosted.nl
ssued	2017-05-12T12:11:36.343Z
Modified	2017-05-12T12:11:36.355Z
icense	license
Catalogs	https://lorentz.fair-dtls.surf-hosted.nl/fdp/catalog/wur
	https://lorentz.fair-dtls.surf-hosted.nl/fdp/catalog/DvBC1
	https://lorentz.fair-dtls.surf-hosted.nl/fdp/catalog/PKD_2.0
	https://lorentz.fair-dtls.surf-hosted.nl/fdp/catalog/AGIS
	https://lorentz.fair-dtls.surf-hosted.nl/fdp/catalog/PKD_1.0
	https://lorentz.fair-dtls.surf-hosted.nl/fdp/catalog/Kenya.2017
	https://lorentz.fair-dtls.surf-hosted.nl/fdp/catalog/test_example_number_2_2.0
	https://lorentz.fair-dtls.surf-hosted.nl/fdp/catalog/post_example_2.0
	https://lorentz.fair-dtls.surf-hosted.nl/fdp/catalog/cds-v1
	https://lorentz.fair-dtls.surf-hosted.nl/fdp/catalog/post-test_1.0
	https://lorentz.fair-dtls.surf-hosted.nl/fdp/catalog/testcatalog_1.0
	https://lorentz.fair-dtls.surf-hosted.nl/fdp/catalog/FAIR_in_Africa_0001





inhibitors p65

Google Search

I'm Feeling Lucky

Mark D. Wilkinson CBGP-UPM/INIA, Madrid

 \odot

SEARCH



ReIA/NFkB p65 Inhibitors: Novus Biologicals https://www.novusbio.com/inhibitors/rela-nfkb-p65 -

RelA/NFkB **p65** Inhibitors available through Novus Biologicals. Browse our RelA/ NFkB **p65** Inhibitor catalog backed by our Guarantee+.

Inhibiting NF-κB Activation by Small Molecules As a Therapeutic ... https://www.ncbi.nlm.nih.gov > NCBI > Literature > PubMed Central (PMC) by SC Gupta - 2010 - Cited by 345 - Related articles May 21, 2010 - ... pathway (Fig 3). Table 1. A list of small molecules as inhibitors of NF-kB pathway Blocking NF-κB activation by inhibitors of p65 acetylation.

Suppression of p65 phosphorylation coincides with inhibition of ... - NCBI https://www.ncbi.nlm.nih.gov/pubmed/16163708 by J Hu - 2005 - Cited by 26 - Related articles







Mark D. Wilkinson CBGP-UPM/INIA, Madrid

FAIR META DATA SEARCH

What data scientists spend the most time dr. ascienceReport 2016.00 tibitors of the Find me all know ^{**}ased on those that Human p65 m those that were were for 1900 found i fives. Keep track of the **Non** for each one. If data is license an relevant, bu please provide the contact information for person I need so I can request the data.

5 September 2018

Google Dataset Search Beta

Q

Search for Datasets

Try boston education data or weather site:noaa.gov

5 September 2018

Google Dataset Search Beta

Q

Search for Dat Biosemanticsgroup LUMC

Try boston education data or weather site:noaa.gov

GOOGLE DATA SET SEARCH

Google Structured Data Testing Tool

http://136.243.4.200.8087/fdp/dataset/gene_disease_association

1 <IDOCTYPE html> 2 <html> <head> < --> Latest compiled and minified CSS --> k rel="stylesheet" href="https://stackpath.bootstrapcdn.com/bootstrap/4.1.0/css/bootstrap.min.css" integrity="sha384-9gVQ4dYFwwWSjIDInLEWnxCjeSWFphJiwGPXrljddIhOegiulFwO5gROvFXOdJI4" crossorigin="anonymous"> <title> Gene disease association (LUMC) 10 11 </title> 12 13 <script type="application/ld+json"> 14 {"@graph": [{"@id":"http://136.243.4.200:8087/fdp/dataset/gene_disease_association","@type":"http://schema.org/Dataset","http ://schema.org/creator":{"@id":"http://biosemantics.org"},"http://schema.org/description":"High-throughput experimental methods such as medical sequencing and genome-wide association studies (GWAS) identify increasingly large numbers of potential relations between genetic variants and diseases. Both biological complexity (millions of potential gene-disease associations) and the accelerating rate of data production necessitate computational approaches to prioritize and rationalize potential gene-disease relations. Here, we use concept profile technology to expose from the biomedical literature both explicitly stated gene-disease relations (the explicitome) and a much larger set of implied gene-disease associations (the implicitome). Implicit relations are largely unknown to, or are even unintended by the original authors, but they vastly extend the reach of existing biomedical knowledge for identification and interpretation of gene-disease associations. The implicitome can be used in conjunction with experimental data resources to rationalize both known and novel associations. We demonstrate the usefulness of the implicitome by rationalizing known and novel gene-disease associations, including those from GNAS. To facilitate the re-use of implicit gene-disease associations, we publish our data in compliance with FAIR Data Publishing recommendations [https://www.forcell.org/group/fairgroup] using nanopublications. An online tool (http://knowledge.bio) is available to explore established and potential gene-disease associations in the context of other biomedical relations.", "http://schema.org/distribution": [{"@id":"http://136.243.4.200:8087/fdp/distribution/gene_disease_association_html"}, {"#id":"http://136.243.4.200:8087/fdp/distribution/gene_disease_association_nguads_gzip"}, {"@id":"http://136.243.4.200:8087/fdp/distribution/gene disease association csv gzip"}],"http://schema.org/keyword s":["The Explicitome", "The Implicitome", "Text mining", "Gene disease association (LUNC)", "GDA", "LNAS"], "http://schema.org/name": "Gene disease association (LUNC)"}, {"#id":"http://biosemantics.org", "#type":"http://schema.org/Thing", "http://schema.org/name": "Biosemantic group"}], "@context":{"rdf":"http://www.w3.org/1999/02/22-rdf-syntax-ns#", "rdfs":"http://www.w3.org/2000/01/rdfschema#", "dcat": "http://www.w3.org/ns/dcat#", "xsd": "http://www.w3.org/2001/XMLSchema#", "owl": "http://www.w3.org/20 02/07/owl#","dcterms":"http://purl.org/dc/terms/","fdp":"http://rdf.biosemantics.org/ontologies/fdpof", "r3d": "http://www.re3data.org/schema/3-0#", "lang": "http://id.loc.gov/vocabulary/iso639-1/"}} 15 </script> 16 17 <style> 18 /* Sticky footer styles 19 •/ 20 html { 21 position: relative; 22 min-height: 100%; 23 24 body { 25 /* Margin bottom by footer height */ 26 margin-bottom: 60px; 27

III O 💽

NEW TEST 2

	0 ERRORS 0 WARNINGS
: http://136.243.4.200:8087/fdp/dataset/gene_disease_association	Dataset
Grype	Dataset
description	High-throughput experimental methods such as medical sequencing and genome-wide association studies (GWAS) identify increasingly large numbers of potential relations between genetic variants and disease. Both biological complexity (millions of potential gene-disease associations) and the accelerating rate of data production necessitate computational approaches to prioritize and rationalize potential gene- disease relations. Here, we use concept profile technology to expose from the biomedical literature both explicitly stated gene-disease sociations (the implicitome) and a much larger set of implied gene-disease associations (the implicitome). Implicit relations are largely unknown to, or are even unintended by the original authors, but they vastly extend the reach of existing biomedical knowledge for identification and interpretation of gene disease associations. The implicitome can be used in conjunction with experimental data resources to rationalize both known and novel associations. We demonstrate the usefulness of the implicitome by rationalizing known and novel gene-disease associations, including those from GWAS. To facilitate the re-use of implicit gene-disease associations, we publish our data in compliance with FAR Data Publishing recommendations [https://www.fore11.org/group/larigroup] using nanopublications. An online tool (http://knowledge.bio) is available to explore established and potential gene-disease associations in the context
	of other biomedical relations.
keywords	The Explicitome
keywords	The Implicitome
keywords	Text mining
keywords	Gene disease association (LUMC)
keywords	GDA
keywords	LWAS
name	Gene disease association (LUMC)
Btype	Thing
ßid	http://biosemantics.org/
Dame	Biosemantic group
distribution	
Øtype	DataDownload
@id	http://136.243.4.200:8087/fdp/distribution/gene_disease_association_htm I
distribution	
@type	DataDownload
®id	http://136.243.4.200:8087/fdp/distribution/gene_disease_association_nqu ads_gzip
distribution	
@type	DataDownload
@id	http://136.243.4.200:8087/fdp/distribution/gene_disease_association_csv

FAIR DATA POINTS AND EXISTING REPOSITORIES

Findable:



F2. data are described with rich metadata;

 F3. metadata clearly and explicitly include the identifier of the data it describes;

F4. (meta)data are registered or indexed in a searchable resource;

Accessible:

 A1. (meta)data are retrievable by their identifier using a standardized communications protocol;

A1.1 the protocol is open, free, and universally implementable;

A1.2. the protocol allows for an authentication and authorization procedure, where necessary;

A2. metadata are accessible, even when the data are no longer available;

Interoperable:



I2. (meta)data use vocabularies that follow FAIR principles;

I3. (meta)data include qualified references to other (meta)data;

Reusable:

R1. (meta)data are richly described with a plurality of accurate and relevant attributes;



R1.1. (meta)data are released with a clear and accessible data usage license;

R1.2. (meta)data are associated with detailed provenance;

R1.3. (meta)data meet domain-relevant community standards;

Luiz Bonino, FigShare webinar State of the Open Data 2018

https://figshare.com/articles/Digital_Science_Webinar_The_State_of_Open_Data_2018/7257221