

Comparison of Annotation and Proposal for Annotation Standard

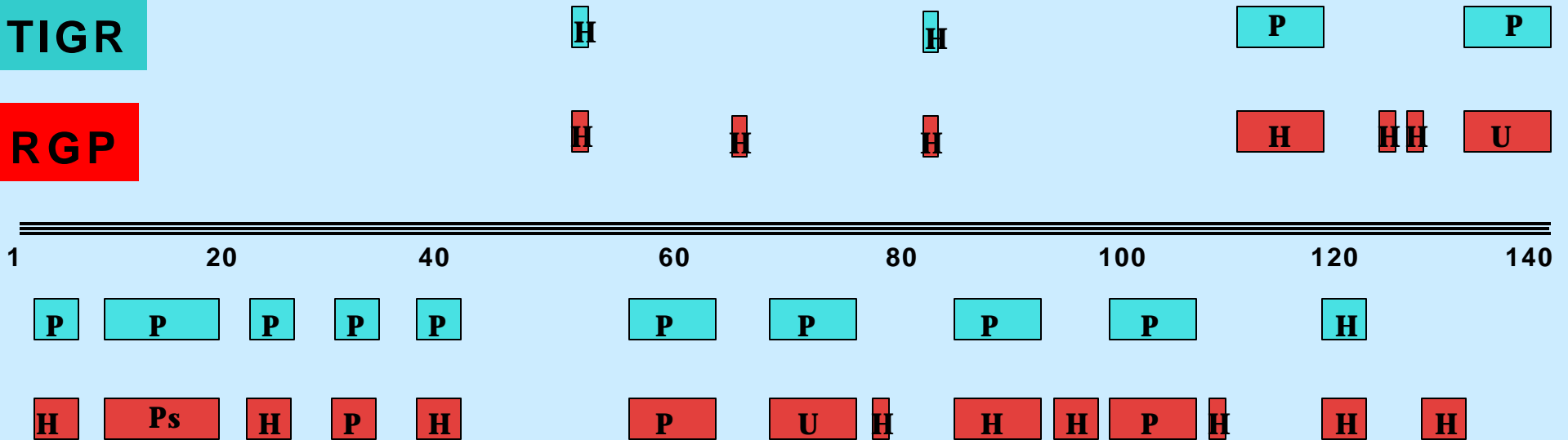
**Katsumi Sakata,
Baltazar A. Antonio, Atsuko Idonuma,
Masatoshi Masukawa, Michie Shibata,
Takashi Matsumoto and Takuji Sasaki**



Example of annotation OSJNBa0067E01 (Chr 3)

TIGR

RGP



RGP

TIGR

Number of predicted genes

21

14

Number of putative genes (P)

3

11

Number of hypothetical genes (H)

15

3

Number of unknown genes (U)

2

0

Summary of annotation

	Predicted		Same		Putative		Hypothetical		Unknown	
	RGP	TIGR	RGP	TIGR	RGP	TIGR	RGP	TIGR	RGP	TIGR
OSJNBa0026O12 (Chr 10)	23	16	0	0	7	8	14	5	2	3
OSJNBa0051D19 (Chr 10)	35	27	0	0	12	14	19	9	4	4
OSJNBa0072E24 (Chr 10)	14	16	0	0	3	9	10	4	1	3
OSJNBa0018H01 (Chr 3)	27	25	0	0	7	15	13	6	6	4
OSJNBa0033N16 (Chr 3)	23	20	1	1	9	12	11	5	2	2
OSJNBa0067E01 (Chr 3)	21	14	0	0	4	11	14	3	2	0
Total	143	118	1	1	42	69	81	32	17	16

~20% more
than TIGR

29 %
of PG

58 %
of PG

Major differences

- 1. Fewer putative and more hypothetical genes predicted by RGP**

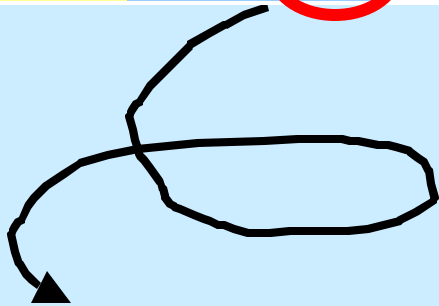
29 % of predicted genes by RGP are putative proteins as compared with 58 % by TIGR.

- 2. Number of predicted genes**

20 % more genes predicted by RGP than TIGR

Difference of nomenclature

RGP	Hypothetical	Hypothetical	Hypothetical	Putative	Putative	Unknown	Unknown		
TIGR	Hypothetical	Putative	Unknown	Putative	Hypothetical	Unknown	Putative		Total
)SJNBa0026O12	0	0	0	3	1	1	0		
)SJNBa0051D19	4	0	0	9	0	3	0	1	
)SJNBb0072E24	1	4	0	2	0	1	0		
)SJNBa0018H01	1	3	1	5	0	2	3	1	
)SJNBb0033N16	3	1	0	7	0	1	1	1	
)SJNBa0067E01	1	4	0	3	0	0	0		
Total	10	12	1	29	1	8	4	6	



Predicted genes with
completely same ORFs
between TIGR and RGP

Investigate to reduce inconsistencies !!!

Highest ranked homology to known protein in BLASTP

Predicted gene name (by RGP)	Bit score	E value	Identity (%)	Ranking in BLASTP	
OSJNBb0072E24.04	201	1.00E-50	41	4	B
OSJNBb0072E24.06	66	5.00E-10	25	2	
OSJNBb0072E24.08	122	2.00E-26	33	3	A
OSJNBb0072E24.11	112	4.00E-24	33	4	B
OSJNBa0018H01.02	84	5.00E-15 #	27	7	C
OSJNBa0018H01.21	125	3.00E-28	46	1	A
OSJNBa0018H01.23	58	3.00E-07	24	11	
OSJNBb0033N16.12	124	4.00E-28	59	9	B
OSJNBa0067E01.01	78	2.00E-13 #	32	8	C
OSJNBa0067E01.03	101	1.00E-20	41	7	B
OSJNBa0067E01.05	89	6.00E-17 #	37	9	C
OSJNBa0067E01.16	743	0	53	18	B

Suggestions to reduce inconsistencies :

A : Standardize BLASTP threshold

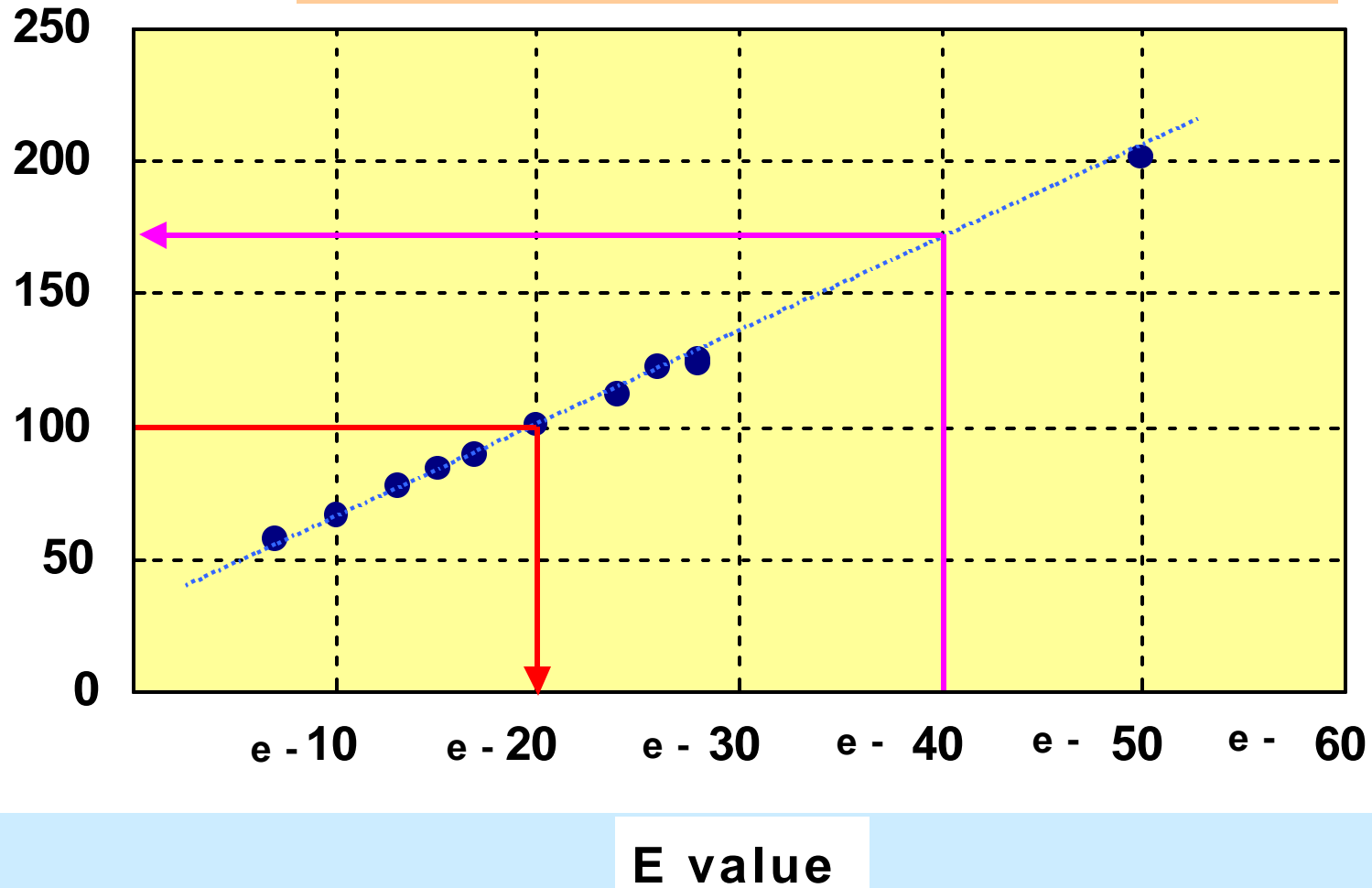
: Higher ranked significant homology to “putative XXX”, “similar to XXX” or “XXX-like protein” exists.

B : Emphasis on homology to a known protein than highly ranked homologies

C : Extend the interpretation of “Putative protein” to include homologies to “putative XXX” etc.

Bit score

Relationship between the bit score and E value of BLASTP



Bit score and E value are in linear relationship

Bit score > 100 is almost equivalent to $E < e^{-20}$

About the number of predicted genes

20 % more genes predicted by RGP than TIGR

We evaluated 30 RGP predicted genes without overlap to TIGR predicted genes.

14 predicted genes are predicted by **multiple** prediction programs.

16 predicted genes are predicted only by **a single** prediction program.

D Standardization of multiple prediction programs will reduce inconsistencies.

Proposed addition and revision to the IRGSP annotation standards (1)

Standard nomenclature for predicted proteins:

Sequences with 100% identity at the amino acid level to known proteins will receive the same, original gene name.

Sequences with less than 100% identity but with significant homology to known proteins will be called "putative" proteins of the same name. The name of the nearest hit will be included as a note. Sequences that are clearly related to a gene family can be called "XXX-like" or "similar to XXX" protein.

Protein matches with BLASTP bit score > 100, E-value < e-20, or other equivalent criteria, will be regarded as significant homologies.

C

A

Proposed addition and revision to the IRGSP annotation standards (2)

Sequences with homology to unknown ESTs will be called "unknown." The EST hit will be included in a note. The homology standard is at least 75% identity at the nucleic acid over ~90% of the length of the entire EST.

Sequences predicted by a **multiple** gene prediction programs with no homology to an EST will be called "hypothetical protein." The gene prediction programs will be included in a note.

Homology to proteins should be well evaluated to estimate the function of as many predicted genes as possible.

D

B