Supplementary Data 1

Table S1. Over-represented abstract terms in the ISG gene list identified using different approaches

This is a four part table showing the results from four different methods: (a) Classical hypergeometric, (b) Permutation test, (c) *Outlier*, and (d) *ExtendedHG*. Over-represented terms were selected by using Bonferroni correction *p*-value ≤ 0.05 as threshold.

Figure S1. Relationship between annotation level and consensus gene age

This is a boxplot showing the number of PMID associated with genes represented on the HG-U133A chip when stratified by the consensus gene age.

Figure S2. Concordance between *Outlier* and *ExtendedHG*

This is a concordance plot showing the rankings for the top 100 most significant tokens found by *Outlier* and *ExtendedHG* when applied to the ISG gene list.

Figure S3. Histograms of hits across different species in *Outlier*

Histogram of the number of tokens identified as over-represented by the *Outlier* method in gene lists derived from experiments performed on 10 Affymetrix platforms. n = number of gene lists available for each platform.

Figure S4. Histograms of hits across different species in *ExtendedHG*

Histogram of the number of tokens identified as over-represented by the *ExtendedHG* method in gene lists derived from experiments performed on 10 Affymetrix platforms. n = number of gene lists available for each platform.

Figure S5. Outlier detection diagnostic plots

This is a 4-panel plot showing the effect of smoothing the local Chip means and SD as a function of the List frequencies using polynomial fitting.

Table S1. Over-represented abstract terms in the ISG gene list identified using different approaches

(a) Classical hypergeometric

Term	Chip	List	*p*-value	Bonferroni *p*-value	Rank
INTERFERON	414	46	1.67E-48	1.96E-44	1
IFN	245	35	2.90E-39	3.40E-35	2
ANTIVIRAL	176	23	5.39E-24	6.33E-20	3
IFN-BETA	71	18	9.52E-24	1.12E-19	4
IFN-ALPHA	114	19	1.38E-21	1.62E-17	5
INDUCIBLE	1068	37	7.02E-20	8.24E-16	6
INTERFERON-ALPHA	59	14	8.13E-18	9.53E-14	7
INFECTION	1177	36	1.60E-17	1.87E-13	8
VIRAL	892	32	3.09E-17	3.62E-13	9
IMMUNE	1275	35	1.58E-15	1.85E-11	10
TREAT	1817	40	6.32E-15	7.41E-11	11
INNATE	363	21	9.16E-15	1.07E-10	12
IFN-GAMMA	443	22	3.43E-14	4.02E-10	13
VIRUS	1408	34	2.07E-13	2.42E-09	14
IMMUNITY	387	20	3.78E-13	4.43E-09	15
DSRNA	60	11	1.25E-12	1.46E-08	16
INDUCTION	2048	39	1.92E-12	2.26E-08	17
OLIGOADENYLATE	18	8	4.64E-12	5.44E-08	18
LYMPHOBLASTOID	239	16	4.80E-12	5.63E-08	19
ISRE	31	9	5.67E-12	6.65E-08	20
HOST	800	24	5.44E-11	6.38E-07	21
ISG	14	7	8.39E-11	9.84E-07	22
MHC	353	17	1.16E-10	1.37E-06	23
TREATMENT	3120	45	1.42E-10	1.66E-06	24
HLA-A	30	8	2.81E-10	3.29E-06	25
STOMATITIS	52	9	4.91E-10	5.76E-06	26
BETA	2127	36	7.56E-10	8.86E-06	27
RESPONSE	3630	47	1.53E-09	1.79E-05	28
HLA-CLASS	11	6	2.43E-09	2.85E-05	29
EVASION	65	9	3.10E-09	3.64E-05	30
CYTOKINE	1266	27	3.37E-09	3.95E-05	31
ANTIGEN	1687	31	3.93E-09	4.60E-05	32
INFECT	825	22	4.13E-09	4.84E-05	33
INDUCE	4669	53	4.27E-09	5.01E-05	34
HLA-B	25	7	4.71E-09	5.52E-05	35
HISTOCOMPATIBILITY	303	14	1.51E-08	0.00018	36
LINE	4667	52	1.62E-08	0.00019	37
HEPATITIS	366	15	1.71E-08	0.00020	38
MELANOMA	581	18	2.22E-08	0.00026	39
ENCEPHALOMYOCARDITIS	16	6	2.25E-08	0.00026	40
REPLICATION	830	21	2.66E-08	0.00031	41
AFTER	3913	46	7.39E-08	0.00087	42
MONOCLONAL	1365	26	7.45E-08	0.00087	43
EPSTEIN-BARR	233	12	8.20E-08	0.00096	44
UPREGULATE	1087	23	1.03E-07	0.00121	45
SYNTHESIS	2200	33	1.28E-07	0.00150	46
BETA2-MICROGLOBULIN	42	7	1.29E-07	0.00151	47
EBV	194	11	1.47E-07	0.00173	48
GAMMA-INTERFERON	44	7	1.72E-07	0.00202	49

Table S1. Over-represented abstract terms in the ISG gene list identified using different approaches (continued)

(a) Classical hypergeometric

Term	Chip	List	*p*-value	Bonferroni *p*-value	Rank
HLA	253	12	1.89E-07	0.00221	50
INTERFERON-GAMMA	313	13	1.95E-07	0.00229	51
OAS	10	5	2.04E-07	0.00239	53
HLA-G	10	5	2.04E-07	0.00239	52
TYPE	4725	50	3.15E-07	0.00369	54
MXA	11	5	3.19E-07	0.00374	55
ALPHA	2422	34	3.50E-07	0.00411	56
DEFINE	2823	37	3.81E-07	0.00447	57
IMMUNODEFICIENCY	472	15	3.95E-07	0.00464	58
PROMYELOCYTIC	216	11	4.00E-07	0.00469	59
INTACT	1382	25	4.03E-07	0.00473	60
LEUKEMIA	1182	23	4.44E-07	0.00521	61
INDEPENDENT	2840	37	4.45E-07	0.00522	62
EACH	3117	39	4.60E-07	0.00539	63
TAPASIN	12	5	4.76E-07	0.00558	64
LYSIS	169	10	4.92E-07	0.00577	65
AUTOIMMUNE	557	16	4.94E-07	0.00580	66
INDIGENOUS	29	6	5.77E-07	0.00677	67
PROTEASOME	490	15	6.20E-07	0.00728	68
LMP2	13	5	6.84E-07	0.00803	69
LMP7	13	5	6.84E-07	0.00803	70
PKR	30	6	6.89E-07	0.00808	71
INDUCIBILITY	131	9	7.86E-07	0.00922	72
CORRESPONDING	2800	36	1.04E-06	0.01214	73
MOLECULE	3217	39	1.07E-06	0.01254	74
DEFENSE	370	13	1.16E-06	0.01362	75
DIFFERENTIAL	1923	29	1.17E-06	0.01368	76
ACTION	1806	28	1.18E-06	0.01386	77
TAP	61	7	1.25E-06	0.01471	78
STIMULATE	2564	34	1.35E-06	0.01581	79
CONFER	1265	23	1.41E-06	0.01651	80
LOAD	383	13	1.67E-06	0.01956	81
REACTIVITY	534	15	1.72E-06	0.02021	82
OR-C	5	4	1.79E-06	0.02103	83
MEDIATE	4505	47	2.06E-06	0.02414	84
RECOMBINANT	2880	36	2.06E-06	0.02420	85
CTL	154	9	2.67E-06	0.03129	86
MICROGLOBULIN	39	6	2.67E-06	0.03137	87
STRAND	1108	21	2.76E-06	0.03238	88
RECOGNIZE	2007	29	2.77E-06	0.03249	89
ALSO	6842	60	3.03E-06	0.03554	90
DERIVE	3496	40	3.12E-06	0.03661	91
P69	6	4	3.57E-06	0.04188	92
VSV	19	5	3.61E-06	0.04240	93
DOUBLE	1235	22	3.76E-06	0.04415	94

Table S1. Over-represented abstract terms in the ISG gene list identified using different approaches (continued)

(b) Permutation

Term	Chip	List	Empirical frequency	Empirical *p*-value	Bonferroni *p*-value
INTERFERON	414	46	0	$< 10^{-5}$	< 0.0484
TREAT	1817	40	0	$< 10^{-5}$	< 0.0484
INDUCIBLE	1068	37	0	$< 10^{-5}$	< 0.0484
IFN	245	35	0	$< 10^{-5}$	< 0.0484
ANTIVIRAL	176	23	0	$< 10^{-5}$	< 0.0484
IFN-ALPHA	114	19	0	$< 10^{-5}$	< 0.0484
IFN-BETA	71	18	0	$< 10^{-5}$	< 0.0484
LYMPHOBLASTOID	239	16	0	$< 10^{-5}$	< 0.0484
INTERFERON-ALPHA	59	14	0	$< 10^{-5}$	< 0.0484
DSRNA	60	11	0	$< 10^{-5}$	< 0.0484
ISRE	31	9	0	$< 10^{-5}$	< 0.0484
EVASION	65	9	0	$< 10^{-5}$	< 0.0484
STOMATITIS	52	9	0	$< 10^{-5}$	< 0.0484
OLIGOADENYLATE	18	8	0	$< 10^{-5}$	< 0.0484
ISG	14	7	0	$< 10^{-5}$	< 0.0484
HLA-CLASS	11	6	0	$< 10^{-5}$	< 0.0484
ENCEPHALOMYOCARDITIS	16	6	0	$< 10^{-5}$	< 0.0484
OR-C	5	4	0	$< 10^{-5}$	< 0.0484
VIRAL	892	32	1	10^{-5}	0.0484
INNATE	363	21	1	10^{-5}	0.0484

100,000 randomisations were performed and the smallest *p*-values that can be measured directly is thus 10^{-5}. 4840 tokens were being tested in this gene list. So the best possible Bonferroni *p*-value attainable is $10^{-5} \times 4840 = 0.0484$. Any term with an empirical *p*-value less than 10^{-5} is provisionally assigned a value of $< 10^{-5}$, and the corresponding Bonferroni *p*-value is deemed to be < 0.0484.

Table S1. Over-represented abstract terms in the ISG gene list identified using different approaches (continued)

(c) ***Outlier***

Term	Chip	List	Z-score	*p*-value	Bonferroni *p*-value	Rank
INTERFERON	414	46	-12.7	2.88E-37	9.81E-34	1
IFN	245	35	-9.64	2.61E-22	8.90E-19	2
IFN-BETA	71	18	-7.59	1.64E-14	5.60E-11	3
ANTIVIRAL	176	23	-6.77	6.56E-12	2.24E-08	4
IFN-ALPHA	114	19	-6.72	8.80E-12	3.00E-08	5
INTERFERON-ALPHA	59	14	-6.62	1.78E-11	6.08E-08	6
OLIGOADENYLATE	18	8	-6.06	6.64E-10	2.26E-06	7
ISG	14	7	-5.76	4.12E-09	1.41E-05	8
ISRE	31	9	-5.70	6.06E-09	2.06E-05	9
DSRNA	60	11	-5.39	3.49E-08	0.00012	10
HLA-CLASS	11	6	-5.29	5.98E-08	0.00020	11
HLA-A	30	8	-5.17	1.16E-07	0.00039	12
HLA-B	25	7	-4.82	7.07E-07	0.00241	13
INDUCIBLE	1068	37	-4.81	7.39E-07	0.00252	14
ENCEPHALOMYOCARDITIS	16	6	-4.74	1.06E-06	0.00361	15
STOMATITIS	52	9	-4.73	1.07E-06	0.00365	16
OAS	10	5	-4.45	4.28E-06	0.01459	17
HLA-G	10	5	-4.45	4.28E-06	0.01459	18
MXA	11	5	-4.33	7.56E-06	0.02577	19
EVASION	65	9	-4.32	7.59E-06	0.02587	20
INNATE	363	21	-4.29	9.13E-06	0.03111	21
TAPASIN	12	5	-4.21	1.25E-05	0.04275	22
VIRAL	892	32	-4.20	1.31E-05	0.04477	23
INFECTION	1177	36	-4.19	1.40E-05	0.04772	24

Table S1. Over-represented abstract terms in the ISG gene list identified using different approaches (continued)

(d) ***ExtendedHG***

Term	**Chip**	**List**	**Odds ratio**	***p*-value**	**Bonferroni *p*-value**	**Rank**
INTERFERON	414	46	2.07	2.24E-35	2.12E-31	1
IFN	245	35	2.01	1.42E-29	1.35E-25	2
IFN-BETA	71	18	2.06	1.41E-18	1.34E-14	3
ANTIVIRAL	176	23	2.02	1.09E-17	1.03E-13	4
IFN-ALPHA	114	19	2.04	2.62E-16	2.49E-12	5
INTERFERON-ALPHA	59	14	2.04	6.17E-14	5.85E-10	6
INDUCIBLE	1068	37	1.99	1.36E-11	1.29E-07	7
VIRAL	892	32	1.98	4.48E-10	4.25E-06	8
OLIGOADENYLATE	18	8	2.15	8.74E-10	8.29E-06	9
INFECTION	1177	36	2.00	1.01E-09	9.61E-06	10
DSRNA	60	11	2.03	1.09E-09	1.03E-05	11
INNATE	363	21	2.01	1.42E-09	1.35E-05	12
ISRE	31	9	2.09	1.71E-09	1.62E-05	13
IFN-GAMMA	443	22	2.00	6.73E-09	6.38E-05	14
ISG	14	7	2.23	9.54E-09	9.05E-05	15
IMMUNITY	387	20	2.01	2.63E-08	0.00025	16
IMMUNE	1275	35	2.01	3.22E-08	0.00030	17
LYMPHOBLASTOID	239	16	2.01	4.58E-08	0.00044	18
HLA-A	30	8	2.15	4.94E-08	0.00047	19
STOMATITIS	52	9	2.09	1.32E-07	0.00125	20
HLA-CLASS	11	6	2.35	1.60E-07	0.00152	21
TREAT	1817	40	2.03	3.50E-07	0.00332	22
HLA-B	25	7	2.23	5.00E-07	0.00475	23
EVASION	65	9	2.09	7.75E-07	0.00735	24
VIRUS	1408	34	1.99	9.37E-07	0.00889	25
MHC	353	17	2.01	1.21E-06	0.0115	26
ENCEPHALOMYOCARDITIS	16	6	2.35	1.44E-06	0.0136	27

Figure S1

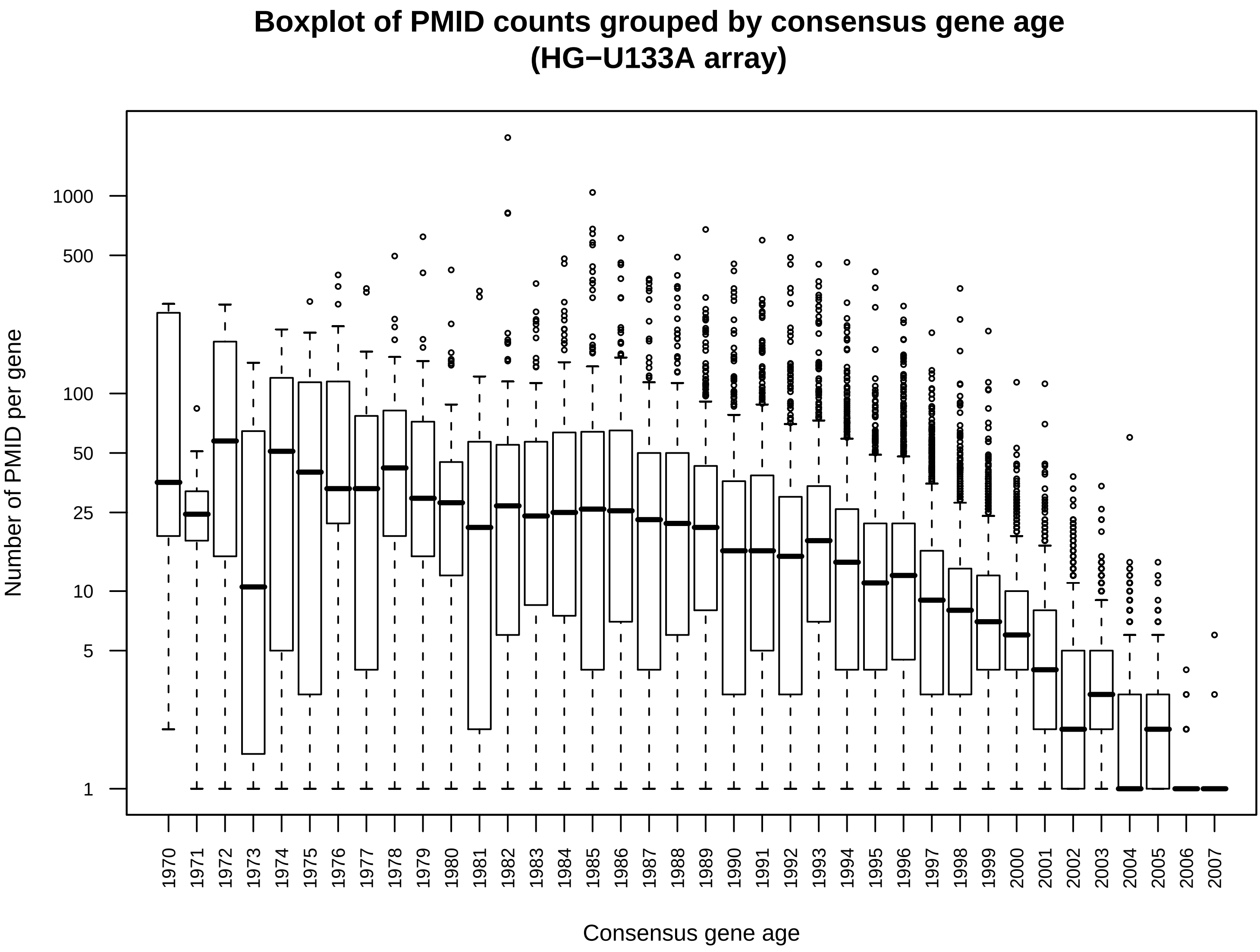
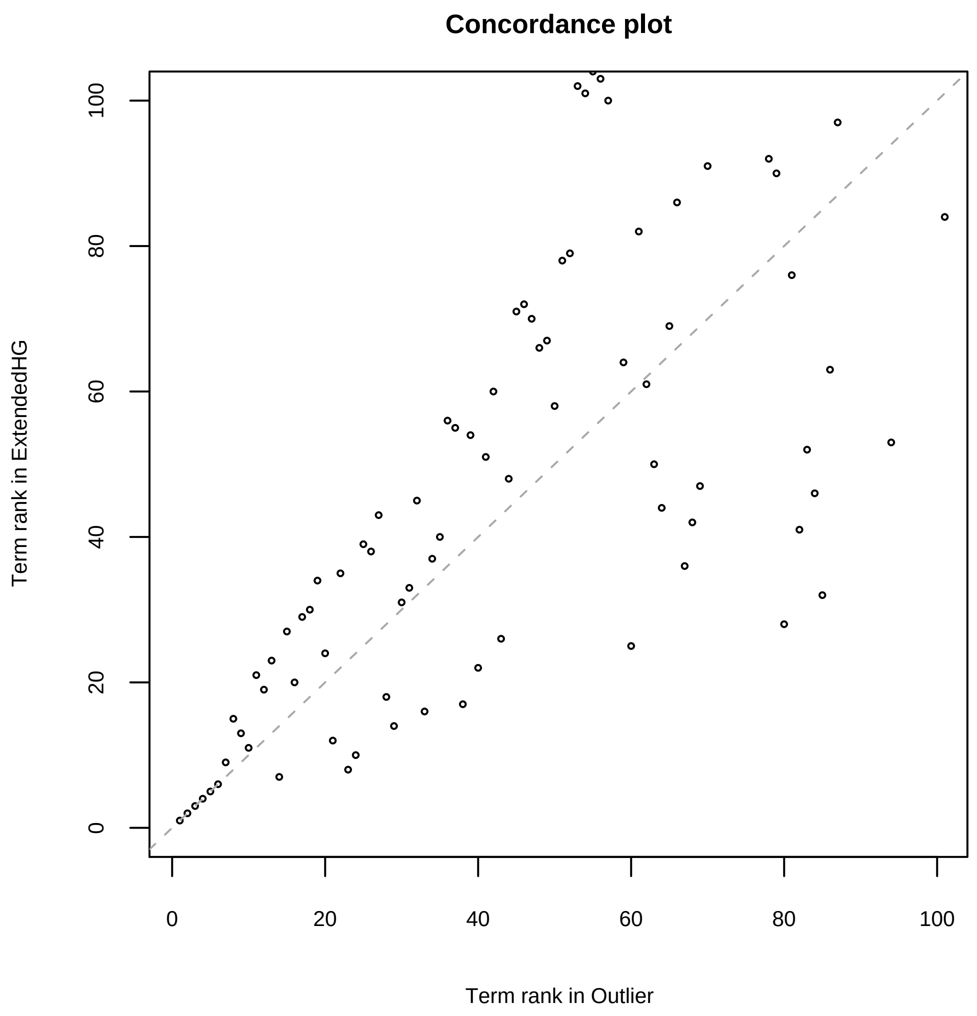

Boxplot of PMID counts grouped by consensus gene age
(HG−U133A array)
Number of PMID per gene
1000
500
100
50
25
10
5
1
1970
1971
1972
1973
1974
1975
1976
1977
1978
1979
1980
1981
1982
1983
1984
1985
1986
1987
1988
1989
1990
1991
1992
1993
1994
1995
1996
1997
1998
1999
2000
2001
2002
2003
2004
2005
2006
2007
Consensus gene age

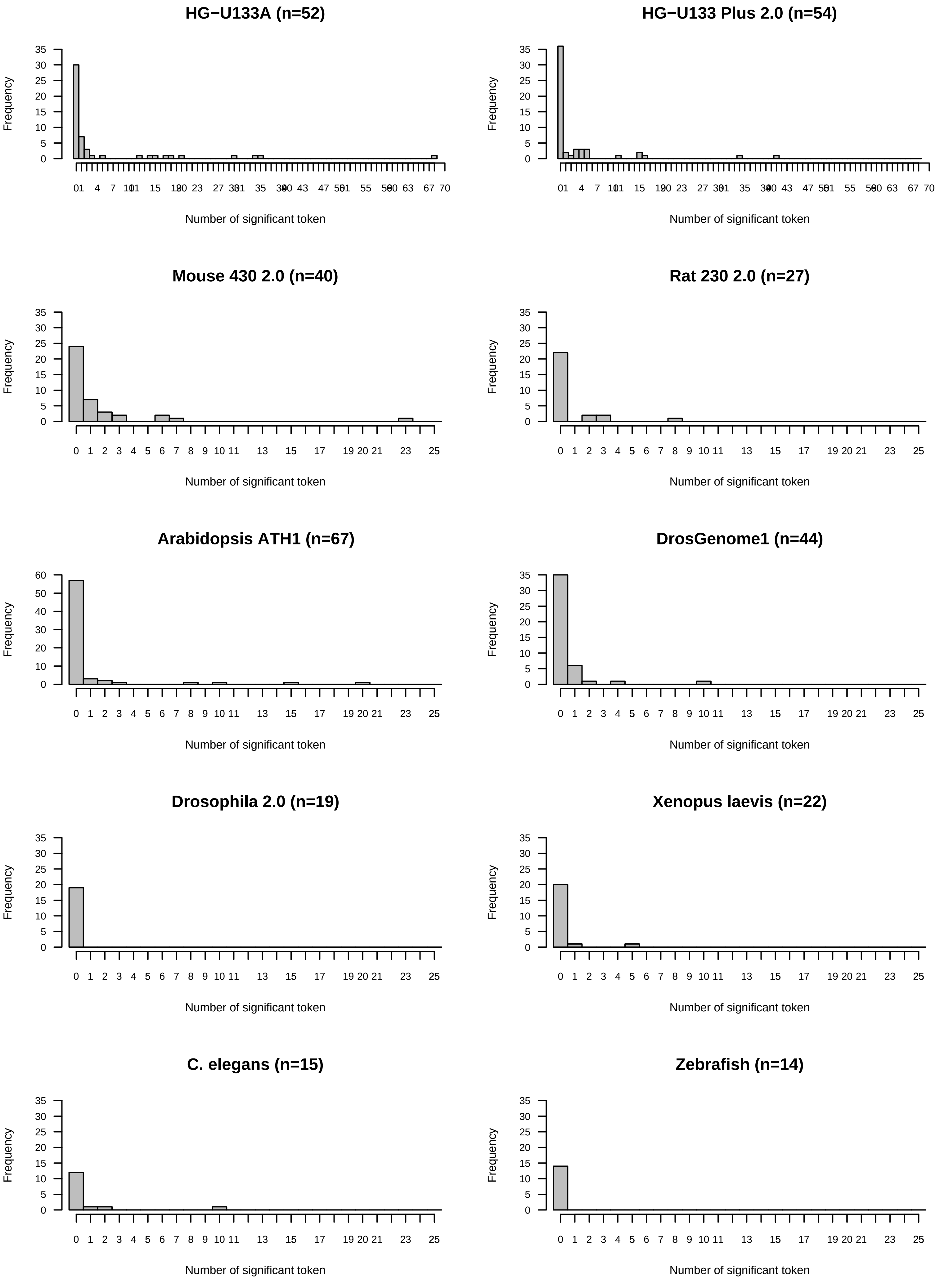
Figure S2



Concordance plot
Term rank in ExtendedHG
0
20
40
60
80
100
Term rank in Outlier
0
20
40
60
80
100

Figure S3

Histograms of hits across different species in Outlier



HG−U133A (n=52)

HG−U133 Plus 2.0 (n=54)

Mouse 430 2.0 (n=40)

Rat 230 2.0 (n=27)

Arabidopsis ATH1 (n=67)

DrosGenome1 (n=44)

Drosophila 2.0 (n=19)

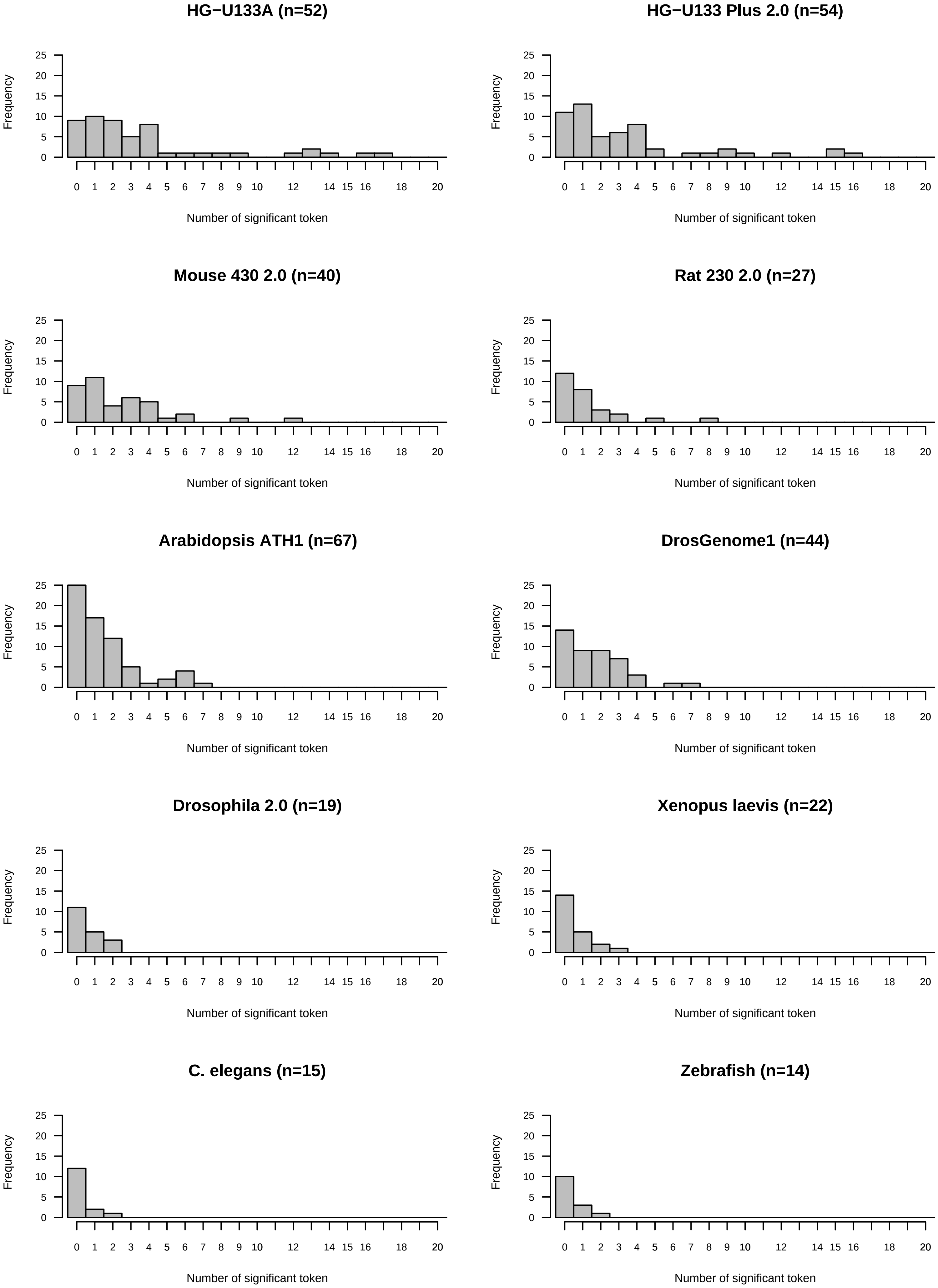
Xenopus laevis (n=22)

C. elegans (n=15)

Zebrafish (n=14)

Figure S4

Histograms of hits across different species in ExtendedHG



HG−U133A (n=52)

HG−U133 Plus 2.0 (n=54)

Mouse 430 2.0 (n=40)

Rat 230 2.0 (n=27)

Arabidopsis ATH1 (n=67)

DrosGenome1 (n=44)

Drosophila 2.0 (n=19)

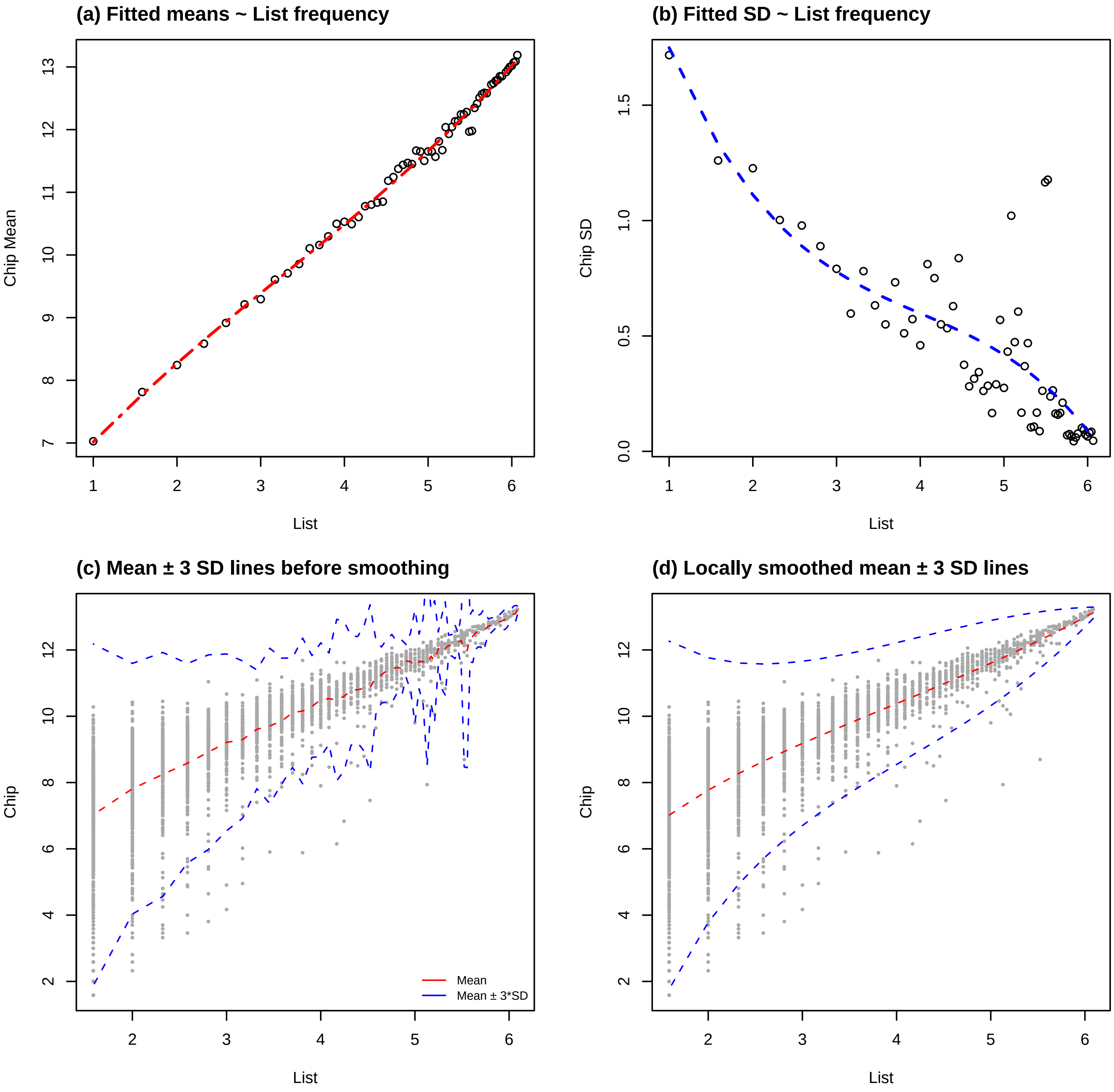
Xenopus laevis (n=22)

C. elegans (n=15)

Zebrafish (n=14)

Figure S5

Outlier detection diagnostic plots



(a) Fitted means ~ List frequency

(b) Fitted SD ~ List frequency

(c) Mean ± 3 SD lines before smoothing

(d) Locally smoothed mean ± 3 SD lines