SUPPLEMENTARY INFORMATION TO
A MATHEMATICAL ALGORITHM FOR DISCOVERING STATES OF EXPRESSION FROM DIRECT GENETIC COMPARISON BY MICROARRAYS

Hassan M Fathallah-Shaykh*, Bin He, Li-Juan Zhao, and Aamir Badruddin

Equation (1.2):

$$f'(x)=\left(\frac{a_1*a_2}{(x+a_2)^2}+\frac{ns+a_3}{(ns-x+a_3)^2}\right)*a_5*\left(\frac{1}{1+\left(\frac{a_7}{x}\right)^{a_6}}+\frac{a_8}{1+\left(\frac{a_{10}}{x}\right)^{a_9}}-\frac{a_{11}}{x+a_{12}}\right)*\left(1+\frac{a_{13}}{1+\left|1-\frac{a_{15}}{x}\right|^{a_{14}}}\right)$$

+

$$\left(\frac{a_1*x}{x+a_2}+\frac{x}{ns-x+a_3}-a_4\right)*a_5*\left(\frac{\frac{a_6}{x}*\left(\frac{a_7}{x}\right)^{a_6}}{\left(1+\left(\frac{a_7}{x}\right)^{a_6}\right)^2}+\frac{a_8*\frac{a_9}{x}*\left(\frac{a_{10}}{x}\right)^{a_9}}{\left(1+\left(\frac{a_{10}}{x}\right)^{a_9}\right)^2}+\frac{a_{11}}{(x+a_{12})^2}\right)*\left(1+\frac{a_{13}}{1+\left|1-\frac{a_{15}}{x}\right|^{a_{14}}}\right)$$

+

$$\left(\frac{a_1*x}{x+a_2}+\frac{x}{ns-x+a_3}-a_4\right)*a_5*\left(\frac{1}{1+\left(\frac{a_7}{x}\right)^{a_6}}+\frac{a_8}{1+\left(\frac{a_{10}}{x}\right)^{a_9}}-\frac{a_{11}}{x+a_{12}}\right)*\left(\frac{a_{13}*a_{14}*sign(a_{15}-x)*\left(\frac{a_{15}}{x^2}\right)*\left|1-\frac{a_{15}}{x}\right|^{a_{14}-1}}{\left(1+\left|1-\frac{a_{15}}{x}\right|^{a_{14}}\right)^2}\right)$$

+

$$\frac{a_{19}*\left(\frac{a_{16}}{x}\right)*\left(\frac{a_{17}}{x}\right)^{a_{16}}}{\left(1+\left(\frac{a_{17}}{x}\right)^{a_{16}}\right)^2}$$

Construction of Equation (1.1) and Legend of Web Figure 1.

Construction of equation (1.1). Our goal in constructing equation (1.1) is to generate a flexible formula that has the ability to adapt in fitting the data sets of the heterogeneous curves of this family. Such an equation is expected to contain parameters that vary to fit differences in curvature at various ranks (x-coordinates). The total number of parameters and complexity are not limiting factors because of the speed of current computers and the availability of mathematical software packages. We start with the equation:

$$y = \frac{x}{ns - x + a_3}$$

ns corresponds to the total number of spots; *ns* = 3840 for the 1.7K microarray. (*a*) shows the effects of varying a_3 on the curve plot. To replicate the rapid elevation of the segment of the curve of (*b*) that corresponds to lower ranks, we add:

$$y = \left(\frac{a_1 * x}{x + a_2} + \frac{x}{3840 - x + a_3} - a_4 \right)$$

(*b*-*d*) show the effects of variations in a_1, a_2, and a_4. Our next goal is to control the slope of the segment of the curve preceding the point of inflection and the position of the inflection rank. (*e*) shows the plot of the equation

$$y = \frac{1}{1+\left(\frac{a_7}{x}\right)^{a_6}}$$

and the effects of varying a_7 and a_6. (*f-g*) plot the equation:

$$y = \left(\frac{a_1 * x}{x + a_2} + \frac{x}{3840 - x + a_3} - a_4\right) * a_5 * \left(\frac{1}{1+\left(\frac{a_7}{x}\right)^{a_6}}\right)$$

and show the effects of variability in a_5, $a_{6,}$, and a_7. (*f*) shows that variations in parameters a_6 and a_7 affect the inflection rank. To gain a fine control on the effects of a_5, a_6 and a_7 on curvature, we added parameters a_8-a_{12}.

$$y = \left(\frac{a_1 * x}{x + a_2} + \frac{x}{3840 - x + a_3} - a_4\right) * a_5 * \left(\frac{1}{1+\left(\frac{a_7}{x}\right)^{a_6}} + \frac{a_8}{1+\left(\frac{a_{10}}{x}\right)^{a_9}} - \frac{a_{11}}{x + a_{12}}\right)$$

Next, we recreate the 'step' in curvature that occurs in association with the inflection point of the curves generated by the human 1.7 and 19K microarrays (see Figure 1 and supplementary information on the web address above). (*h*) shows the plot of the equation:

$$y = 1 + \frac{a_{13}}{1+\left|1-\frac{a_{15}}{x}\right|^{a_{14}}}$$

where a_{15} =1000 and the effects of variations in a_{14}. (*i*) shows the effects of varying a_{14} and a_{15} on the plot of:

$$y = \left(\frac{a_1 * x}{x + a_2} + \frac{x}{3840 - x + a_3} - a_4 \right) * a_5 * \left(\frac{1}{1 + \left(\frac{a_7}{x} \right)^{a_6}} + \frac{a_8}{1 + \left(\frac{a_{10}}{x} \right)^{a_9}} - \frac{a_{11}}{x + a_{12}} \right) * \left(1 + \frac{a_{13}}{1 + \left| 1 - \frac{a_{15}}{x} \right|^{a_{14}}} \right)$$

(*i*) also shows that variations in a_{14} and a_{15} also affect the inflection rank. The curves generated by the lymphoma lack a 'step' at the inflection point; there, parameters a_{14} and a_{15} are applied to recreate the rapid rise in slope seen at higher ranks (see supplementary information). Finally, we added parameters a_{16}-a_{19} in:

$$\left(\frac{1}{1 + \left(\frac{a_{17}}{x} \right)^{a_{16}}} - a_{18} \right) * a_{19}$$

to control the slope of its initial rise (a_{16} and $a_{17,}$ (j)) and the vertical position of the curve (a_{18} and a_{19}, (*k-l*)). Current computers handle Equation (1.1) efficiently.

Legend of Web Figure 2

Equation (1.1) fits the data sets of 60/60 human 1.7K (1.7K), 200/200 human 19K (19KP1 and 19kP2)), and all 266 publicly available curves of the lymphoma study (Alizadeh AA al. *Nature* 2000, 403: 503-511). Negative background-subtracted intensity measurements are transformed to 1 (log = 0). Column *a* shows the log transformation of the background-subtracted intensity measurements in channel 1 (red) and channel 2 (magenta). Column *b* and *d* superimpose the plots of equation 1.1 (black and blue) onto the corresponding data sets in channels 1 and

2 (red and magenta), respectively. Columns *c* and *d* show the plots of equation 1.2 corresponding to the data sets of channels 1 and 2, respectively. The enclosed Excel sheet contains the 266 *(*$a_1,\ldots,a_{19}$) parameters of equation (1.1) and the norms of residuals of the fits.

Normalization And Spot Ranks

Fitting the sorted data of symmetrical images into equation (1.1) generates 2 intensity curves I_1 and I_2 (Figure 1a), 2 log curves L_1 and L_2 (Figure 1b), and 2 separate Spot Orders (SO1 and SO2). Each spot can thus be represented by its 2 ranks $(r_1(s), r_2(s))$ in SO1 and SO2, respectively. Normalization is applied to annul confounding experimental variations in labeling, laser and probe intensities. It is based on the idea that the expression levels of the predominant majority of genomic spots do not differ between 2 samples. Normalization implies that the products of all the expression measurements of the 2 samples are equal; thus:

$$\prod_{s=1}^{3840}\left(\frac{e^{\mathrm{L}_1(r_1(s))}}{e^{\mathrm{L}_2(r_2(s))}}\right)=1 \quad \textbf{(2.1)}$$

Formula **(2.1)** leads to:

$$\frac{\prod_{p=1}^{3840} e^{\mathrm{L}_1(p)}}{\prod_{p=1}^{3840} e^{\mathrm{L}_2(p)}}=1, \text{ and } \sum_{p=1}^{3840}(\mathrm{L}_1(p)-\mathrm{L}_2(p))=0 \quad \textbf{(2.2)}$$

Where p denotes the same rank in SO1 and SO2. Formula (2.2) uncovers a new strategy for normalization (Figure 1e). Specifically, the y-coordinates or expression values of the spots of SO2 are transformed to become equal to the expression values of spots having equal ranks in SO1. For instance, to normalize the

expression data of L_2, shown in Figure 1d, to model the log curve L_1, shown in Figure 1b, a normalization function N is applied to the data of L_2 such as:

$$N(p) = I_1(p) \quad \text{and}$$

$$\text{normalized } L_2(p) = \log(N(p)) = L_1(p) \quad \textbf{(2.3)}$$

Where p denotes the same rank in SO1 and SO2. A specific rank may correspond to 2 different spots in SO2 and SO1. Hence, the normalized $L_2(p)$ satisfies equation (2.2).

Let g be the normalized function, and let *a* and *b* be the ranks of a single spots in symmetrical images, then the

$$\text{Normalized ratio} = \frac{e^{g(a)}}{e^{g(b)}} = e^{g(a)-g(b)}$$

RT-PCR Primers/Annealing Temperatures:

G3PDH
Forward primer: 5'-CAAGGTCATCCCTGAGCTGAAC-3'
Reverse primer: 5'-TCGCTGTTGAAGTCAGAGGAGAC-3'
Annealing Temperature: 60^{O}C

SN25
Forward primer: 5'-CAAGTTGGCTGATGAGTCGCTG-3'
Reverse primer: 5'-TGATTTGGTCCATCCCTTCCTC-3'
Annealing Temperature: 60^{O}C

M6A
Forward primer: 5'-AATGCTGTATCAAATGCCTGGG-3'
Reverse primer: 5'-GCCATCTCAAATGTAGGTTTGCAG-3'
Annealing Temperature: 60^{O}C

143F
Forward primer: 5'-CCTACAAGAATGTGGTTGGTGCC-3'
Reverse primer: 5'-GCAAACTGTCTCCAGCTCCTTCTC-3'
Annealing Temperature: 60^{O}C

CHIN
Forward primer: 5’-CCATCCAAAGAGTCTTGGTCAGG-3’
Reverse primer: 5’-GTTGCCAGATTGTCACAGACGAC-3’
Annealing Temperature: 60^{O}C

CALM
Forward primer: 5’-TGGCTGACCAACTGACTGAAGAG-3’
Reverse primer: 5’-GTAACTCTGCTTCTGTGGGATTCTG-3’
Annealing Temperature: 60^{O}C

MYPR
Forward primer: 5’-CTCAAAGGGTACTTCCACTGATGG-3’
Reverse primer: 5’-CAATAGGCAGATTTGGGCAAAC-3’
Annealing Temperature: 60^{O}C

TBB3
Forward primer: 5’-ATGGGCACGTTGCTCATCAG-3’
Reverse primer: 5’-TCGTTGTCGATGCAGTAGGTCTC-3’
Annealing Temperature: 60^{O}C

HB2J
Forward primer: 5’-TGGTCTGCTCTGTGAGTGGTTTC-3’
Reverse primer: 5’-TGTTTCCAGCATCACCAGGGTC-3’
Annealing Temperature: 60^{O}C

K2C8
Forward primer: 5’-CCATTAAGGATGCCAACGCC-3’
Reverse primer: 5’-GACGTTCATCAGCTCCTGGTACTC-3’
Annealing Temperature: 60^{O}C

MT2
Forward primer: 5’-GCAAATGCACCTCCTGCAAG-3’
Reverse primer: 5’-CGTTCTTTACATCTGGGAGCGG-3’
Annealing Temperature: 60^{O}C

RLA1
Forward primer: 5’-CATTCTGCACGACGATGAGGTG-3’
Reverse primer: 5’-CAGATGAGGCTCCCAATGTTGAC-3’
Annealing Temperature: 60^{O}C

PTPZ
Forward primer: 5’-AAACCTCGTGGAGAAAGGAAGG-3’
Reverse primer: 5’-GCGTGTAGTGATACTGTGTGACCAC-3’
Annealing Temperature: 60^{O}C

EAT1
Forward primer: 5'-TGTGAGGACAGACAGAAGGCAAAG-3'
Reverse primer: 5'-TGGGCTGGAAGAGACCATGAAGAC-3'
Annealing Temperature: 60^{O}C

TBB5
Forward primer: 5'-GAGTTCCCAGACCGCATCATG-3'
Reverse primer: 5'-GATGTCGTAGAGTGCCTCGTTGTC-3'
Annealing Temperature: 60^{O}C

OSTP
Forward primer: 5'-TGCAGCCTTCTCAGCCAAAC-3'
Reverse primer: 5'-CCACAGCATCTGGGTATTTGTTG-3'
Annealing Temperature: 60^{O}C

AMO2
Forward primer: 5'-TGTGCCAGGACTCTCTTTCTTCC-3'
Reverse primer: 5'-AAGGTTCAGTGTCCCCTGTGTCAG-3'
Annealing Temperature: 60^{O}C

PTRR
Forward primer: 5'-TCAAGGACGCTGTGCTCTACTCTG-3'
Reverse primer: 5'-CCAGGAAGTAAAGGAAGAAGGTCAC-3'
Annealing Temperature: 60^{O}C

SPCO
Forward primer: 5'-TCAGGGTGTTTGGCATGTCC-3'
Reverse primer: 5'-TGTGGGAGGAGATGAAGACCAC-3'
Annealing Temperature: 60^{O}C

ANX5
Forward primer: 5'-CCCTCATCAGAGTCATGGTTTCC-3'
Reverse primer: 5'-TCATCTTCTCCACAGAGCAGCAG-3'
Annealing Temperature: 58^{O}C

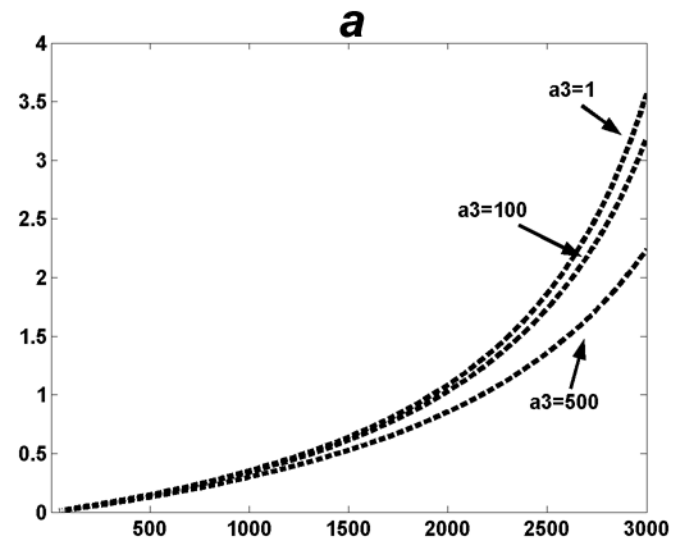
RL10
Forward primer: 5'-AATTCGGCATGAGGGACCAC-3'
Reverse primer: 5'-AATCTTGGCATCAGGGACACC-3',
Annealing Temperature: 56^{O}C

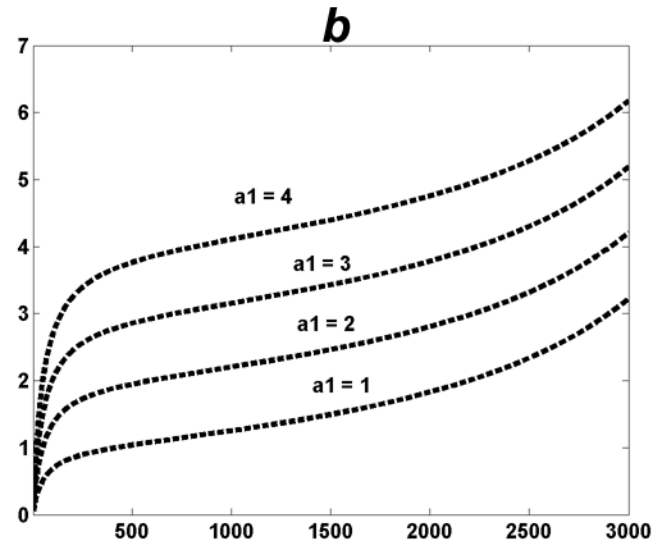
MAPB
Forward primer: 5'-CGCTGAACAATTACCTGCCAAATG-3'
Reverse primer: 5'-CCTAGCAGAAGTCAGTTGTGTTGG-3'
Annealing Temperature: 60^{O}C

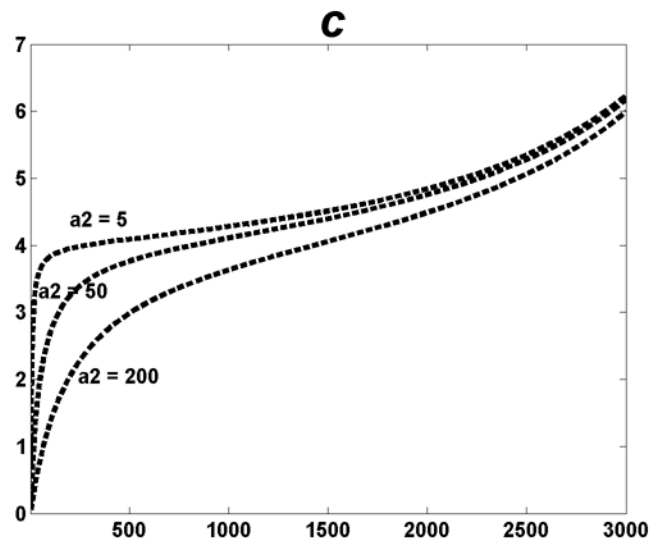
Legend of Web Figure 3

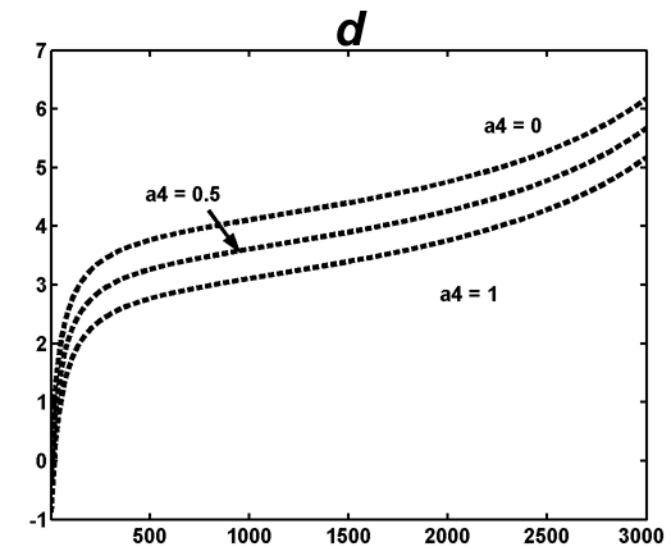
The algorithm is applied to analyze the same-to-same (split-control) data of Rosenzweig et al. (Environmental Health Perspectives 112:480-487, 2004). H001451, H001476, H001488, H001494 are the symmetrical sets (dye swap) of H001452, H001477, H001489 H001495, respectively. The Figure plots the raw data (black and magenta) and curve fits (red and blue) of the Cy3 (black and red) and Cy5 (magenta and blue) images. The green arrows point to the CR.

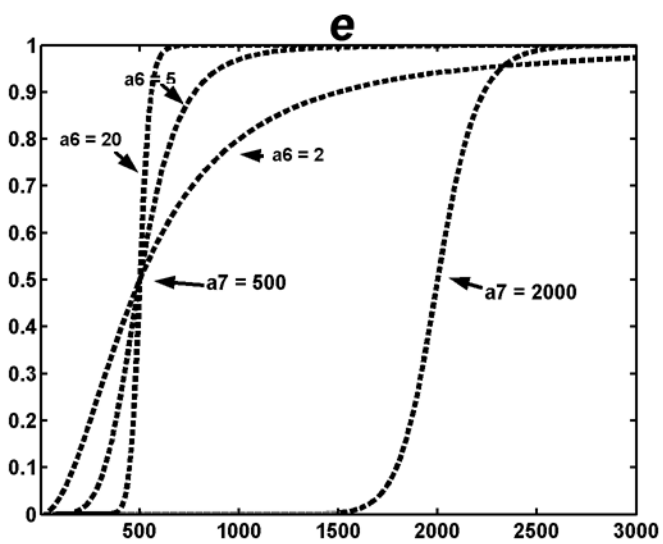
Web Figure 1

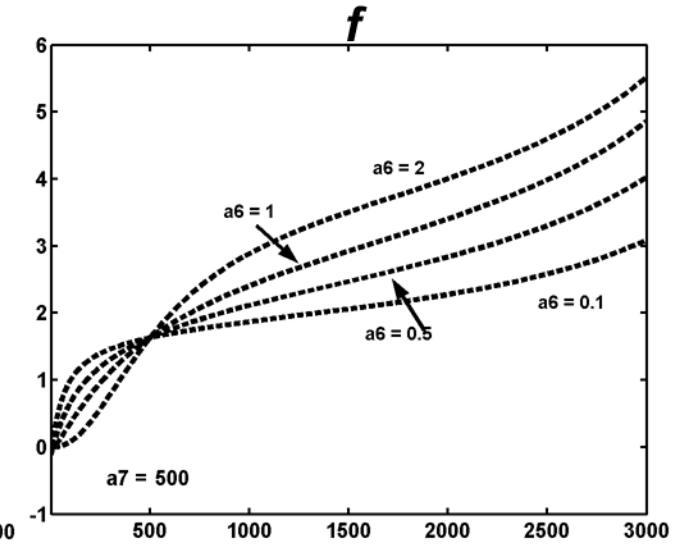

a
a3=1
a3=100
a3=500

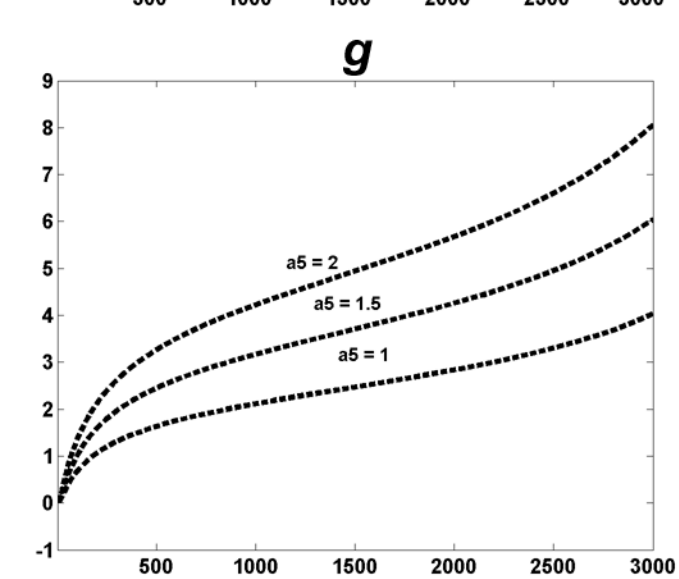

b
a1 = 4
a1 = 3
a1 = 2
a1 = 1

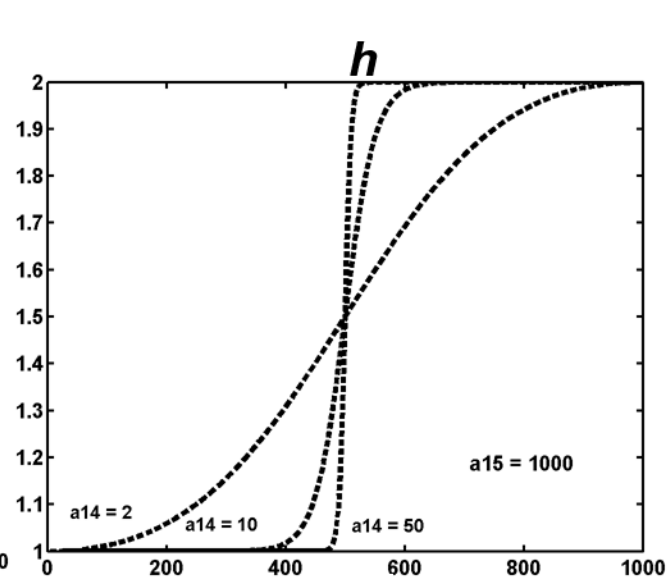

c
a2 = 5
a2 = 50
a2 = 200

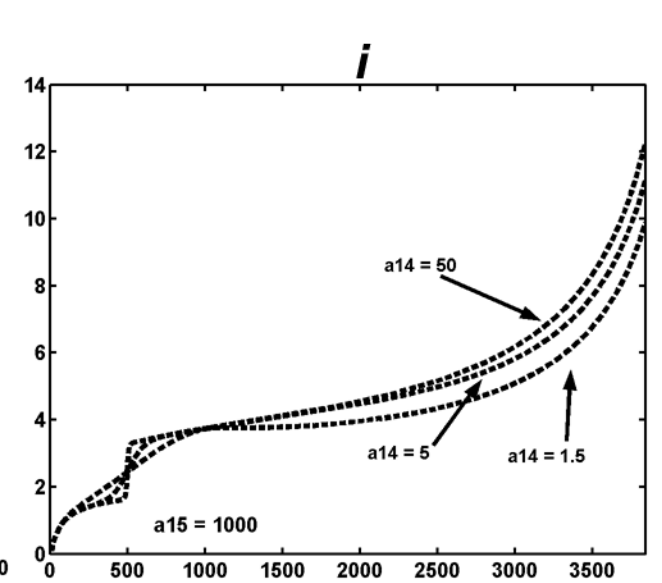

d
a4 = 0
a4 = 0.5
a4 = 1

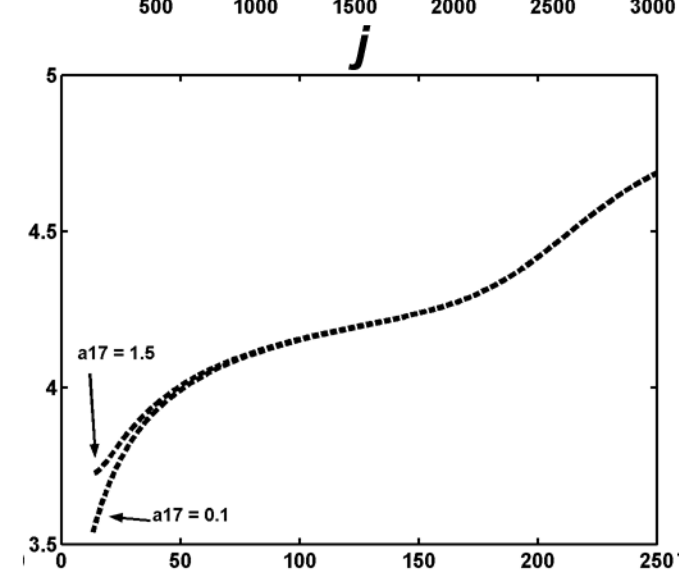

e
a6 = 5
a6 = 20
a6 = 2
a7 = 500
a7 = 2000

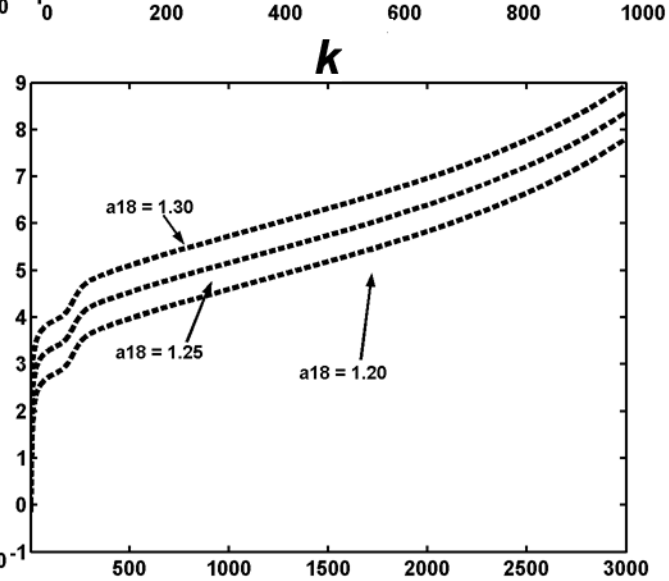

f
a6 = 2
a6 = 1
a6 = 0.1
a6 = 0.5
a7 = 500

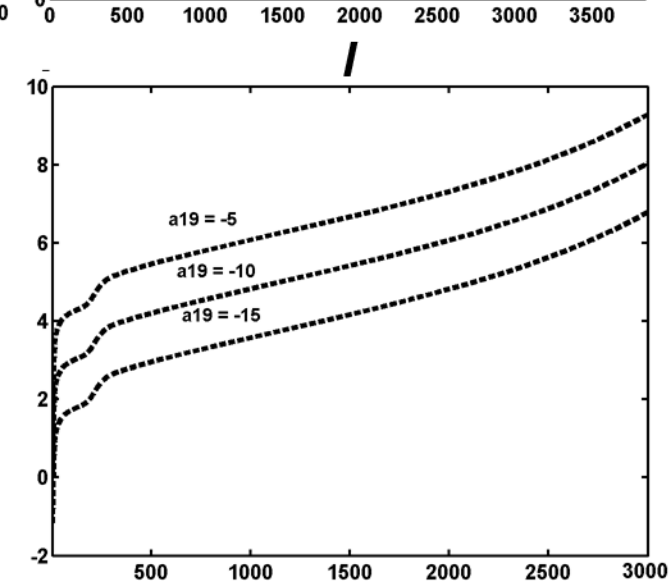

g
a5 = 2
a5 = 1.5
a5 = 1


h
a14 = 2
a14 = 10
a14 = 50
a15 = 1000

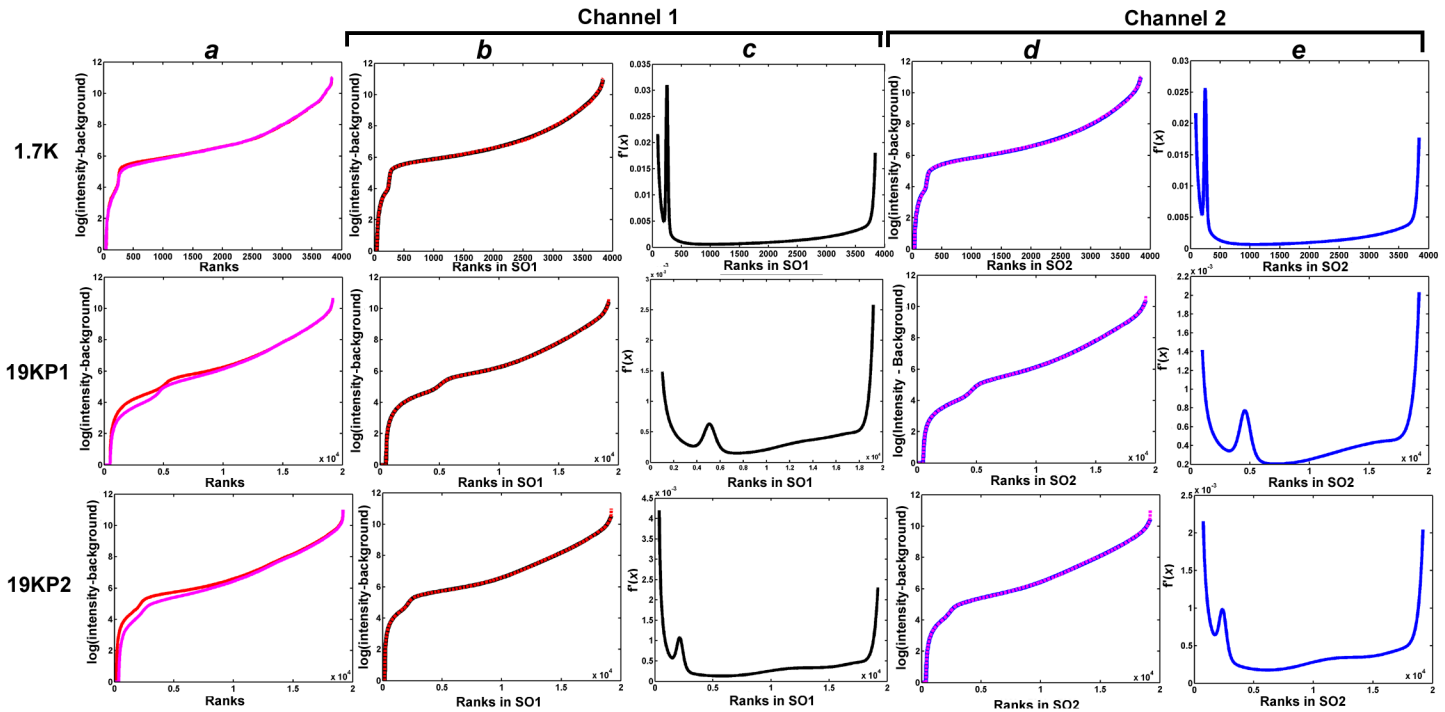

i
a14 = 50
a14 = 5
a14 = 1.5
a15 = 1000


j
a17 = 1.5
a17 = 0.1

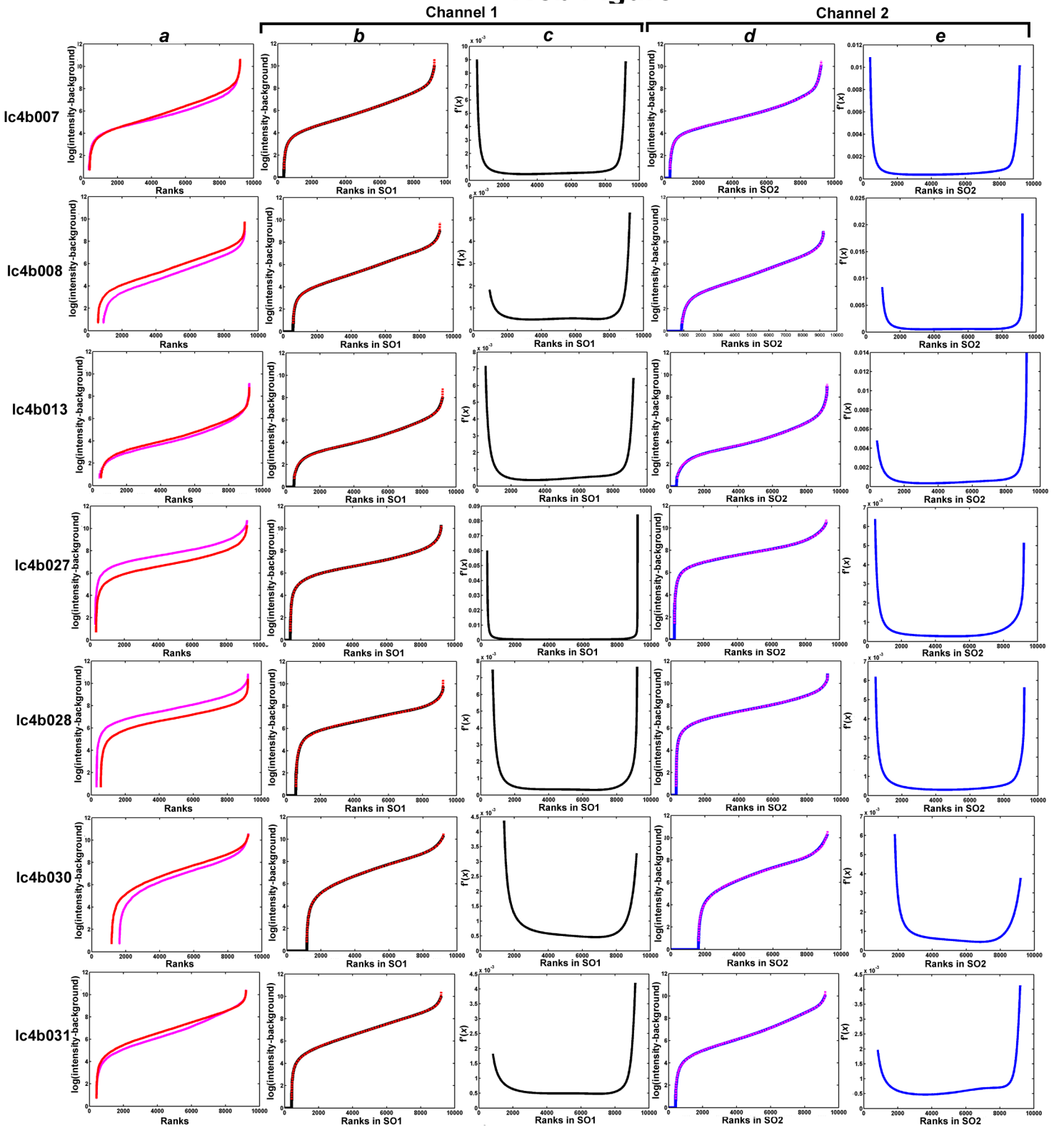

k
a18 = 1.30
a18 = 1.25
a18 = 1.20


l
a19 = -5
a19 = -10
a19 = -15

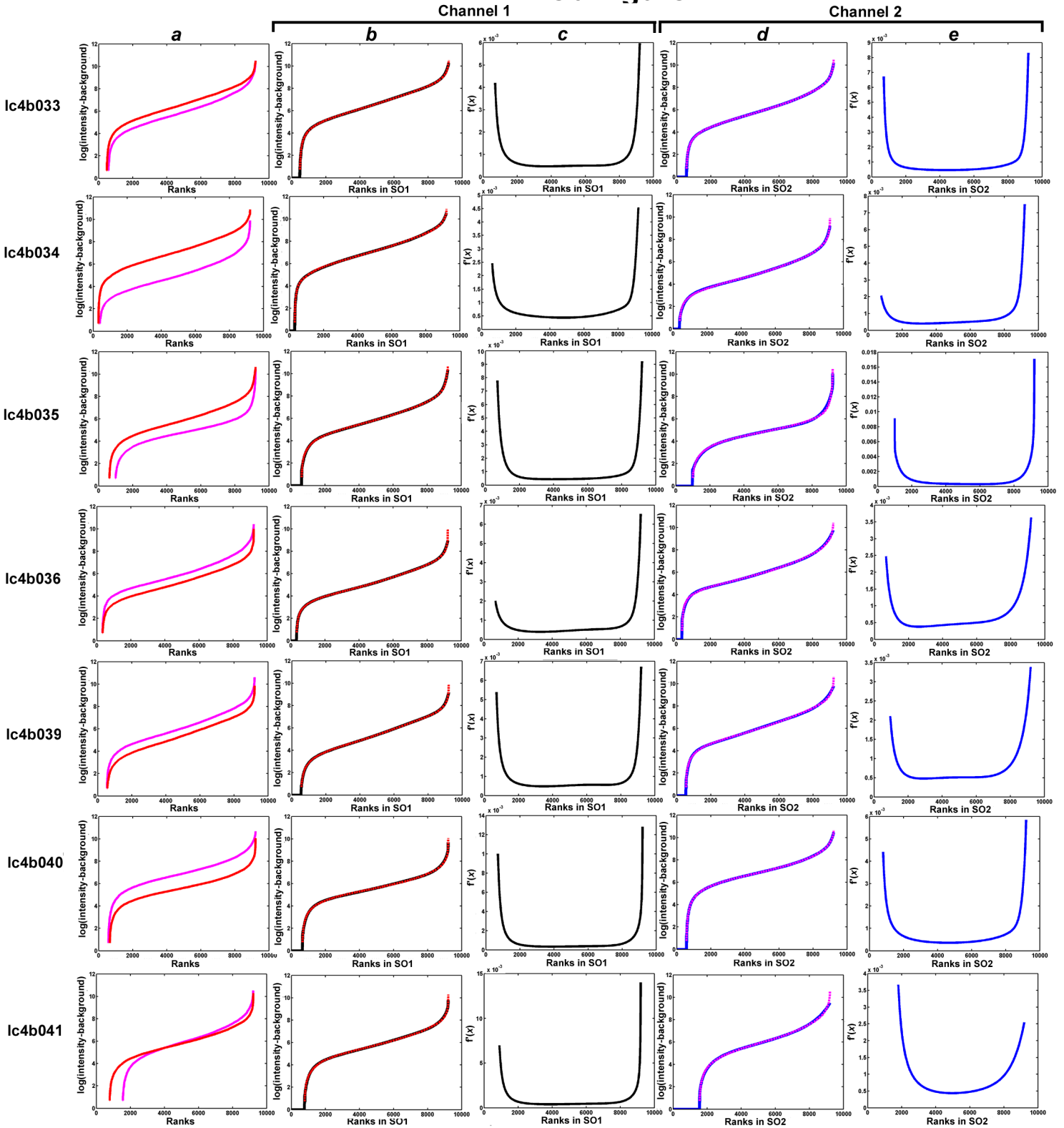
Web Figure 2


Channel 1
Channel 2
a
b
c
d
e
1.7K
19KP1
19KP2
log(intensity-background)
log(Intensity - Background)
f'(x)
Ranks
Ranks in SO1
Ranks in SO2

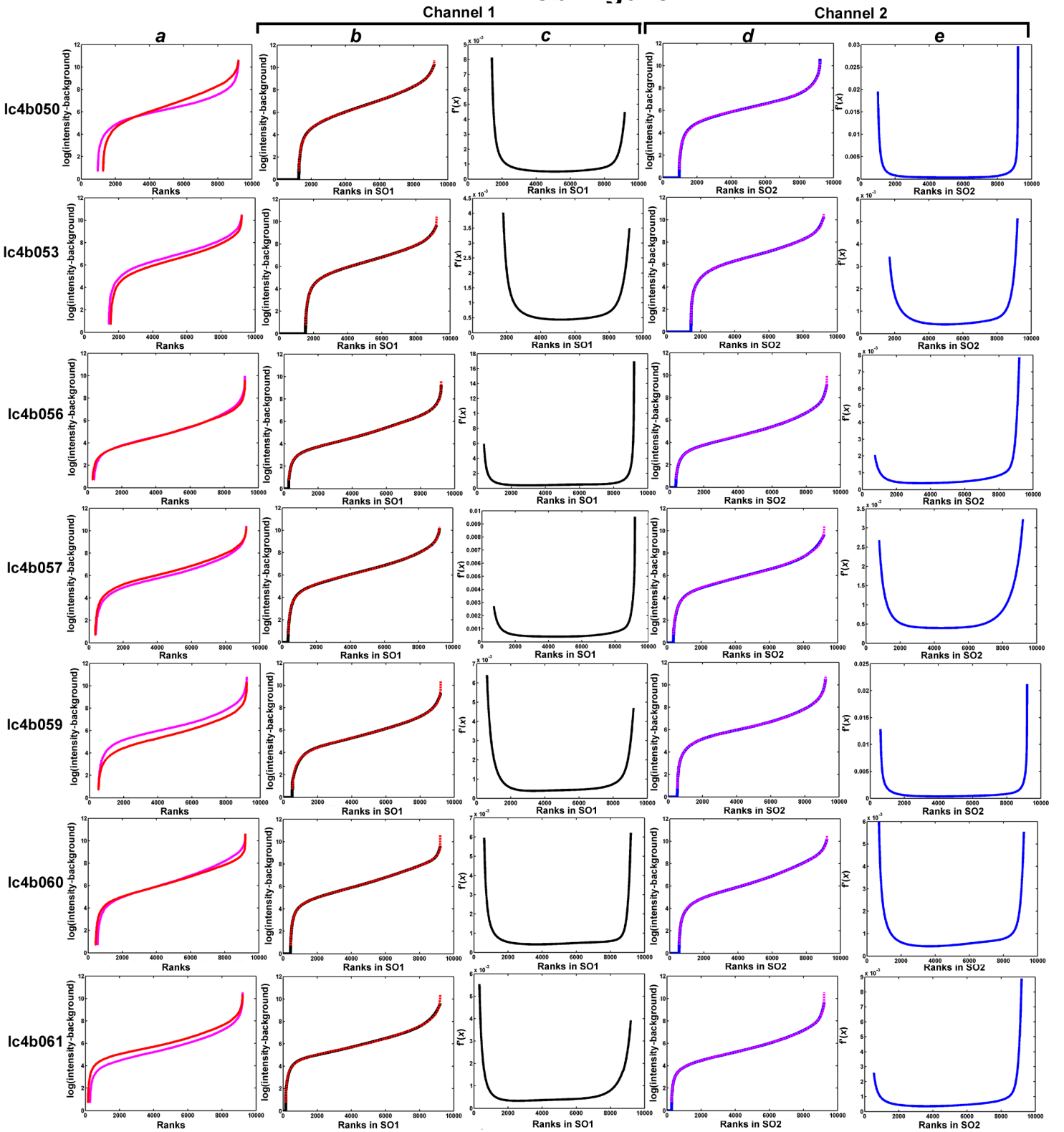
Web Figure


Channel 1
Channel 2
a
b
c
d
e
lc4b007
lc4b008
lc4b013
lc4b027
lc4b028
lc4b030
lc4b031
log(intensity-background)
f'(x)
Ranks
Ranks in SO1
Ranks in SO2

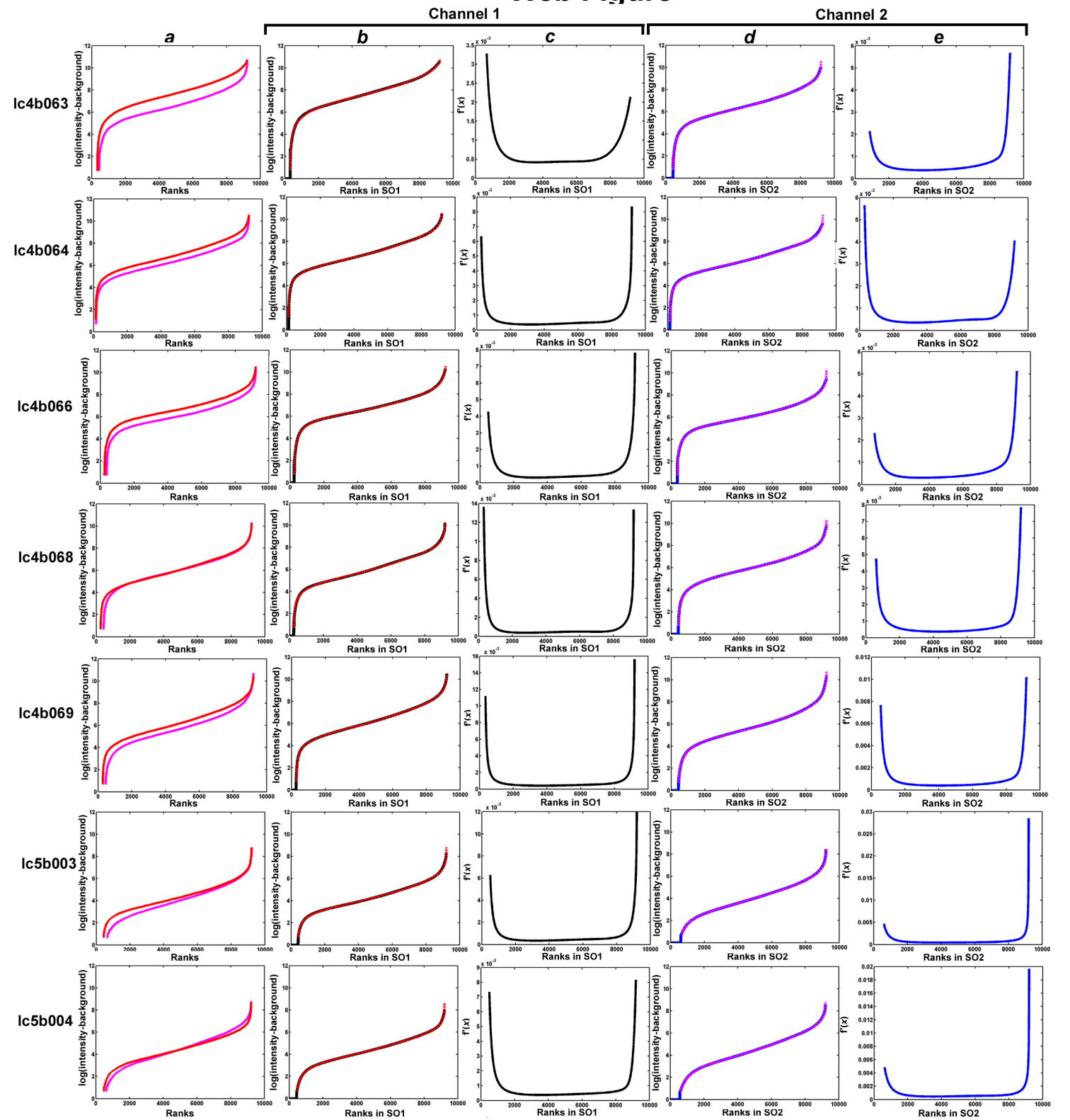
Web Figure


Channel 1
Channel 2
a
b
c
d
e
lc4b033
lc4b034
lc4b035
lc4b036
lc4b039
lc4b040
lc4b041
log(intensity-background)
f'(x)
Ranks
Ranks in SO1
Ranks in SO2

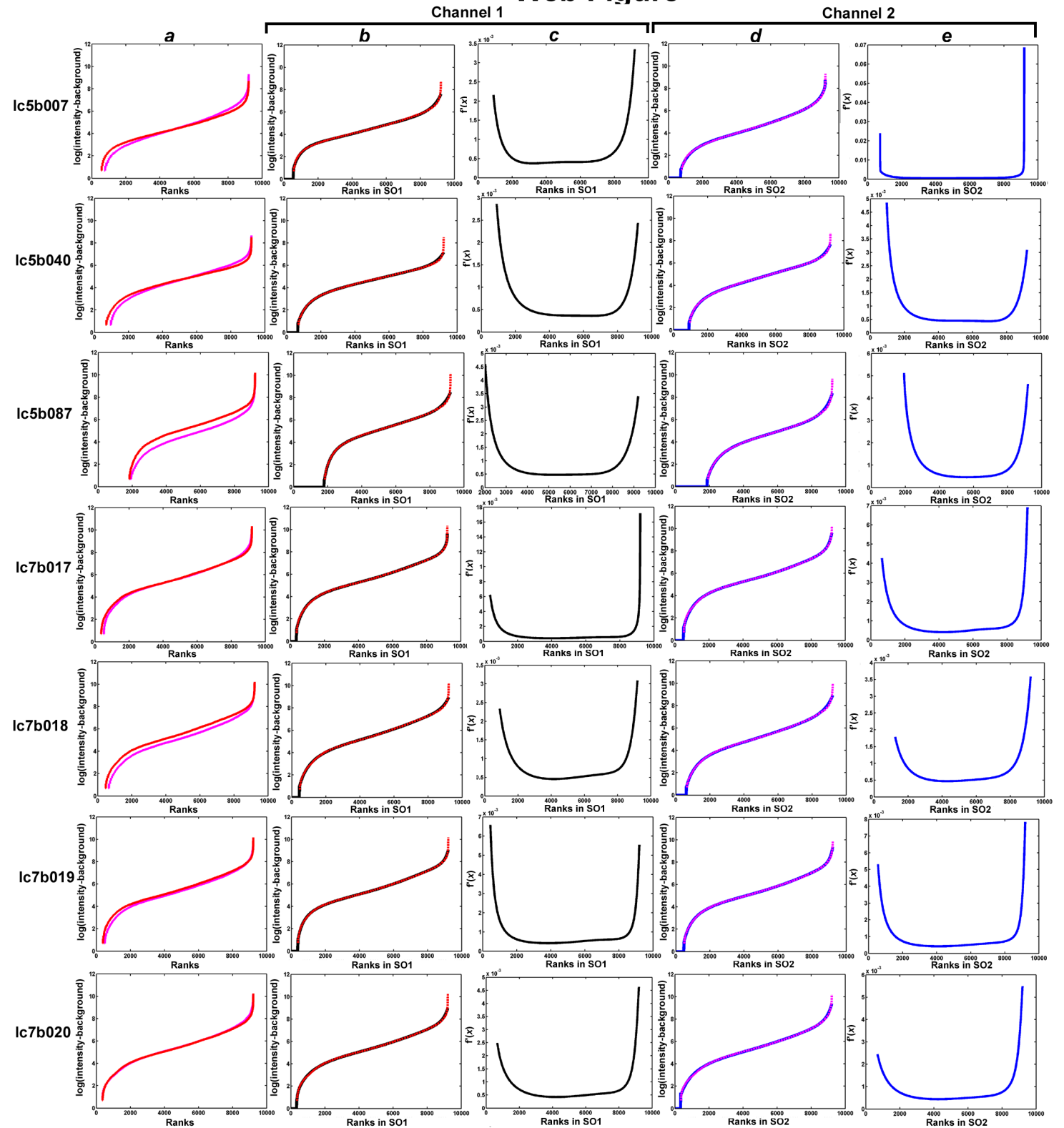
Web Figure


Channel 1
Channel 2
a
b
c
d
e
lc4b050
lc4b053
lc4b056
lc4b057
lc4b059
lc4b060
lc4b061
log(intensity-background)
f'(x)
Ranks
Ranks in SO1
Ranks in SO2

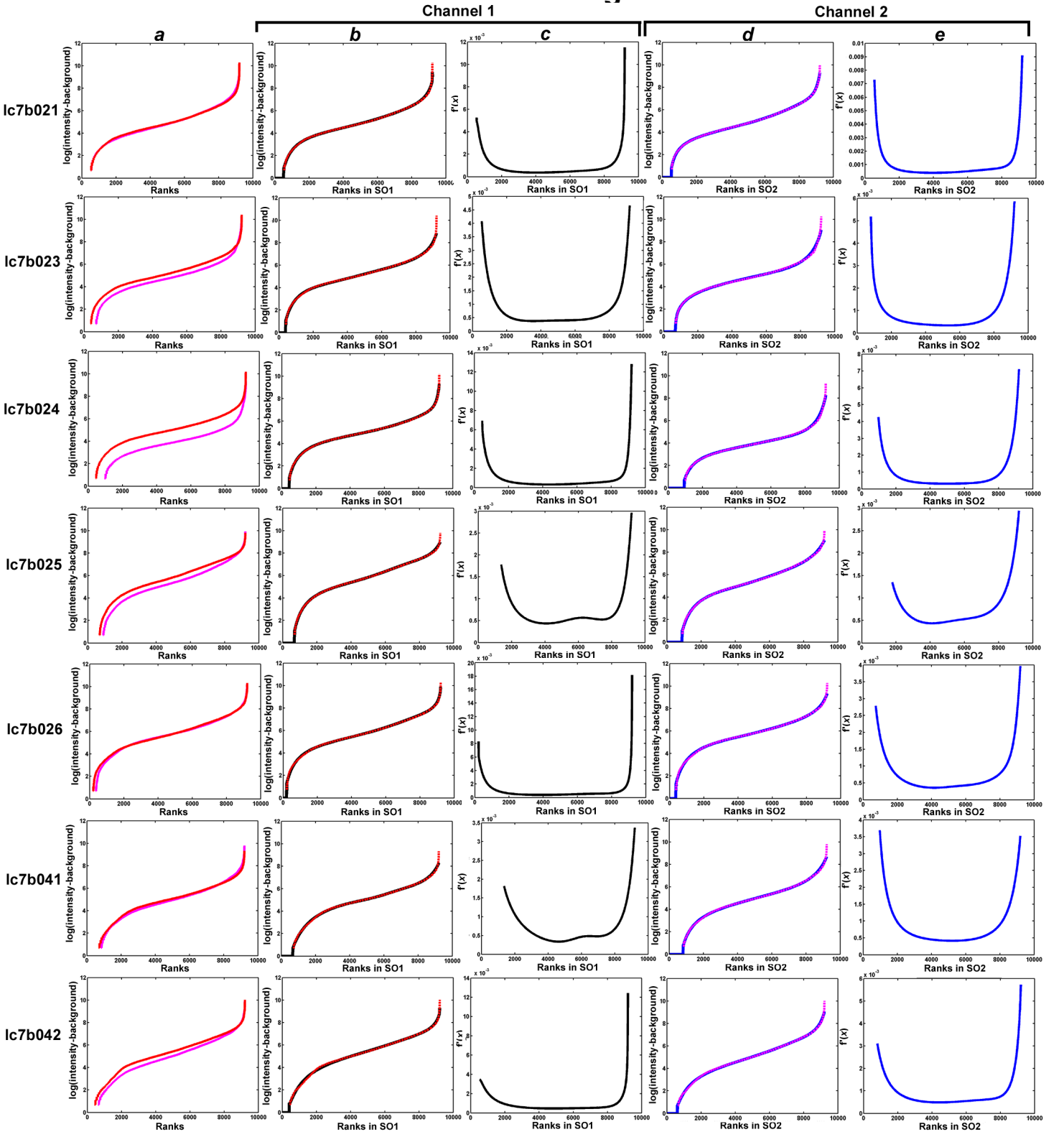
Web Figure


Channel 1
Channel 2
a
b
c
d
e
lc4b063
lc4b064
lc4b066
lc4b068
lc4b069
lc5b003
lc5b004
log(intensity-background)
f'(x)
Ranks
Ranks in SO1
Ranks in SO2

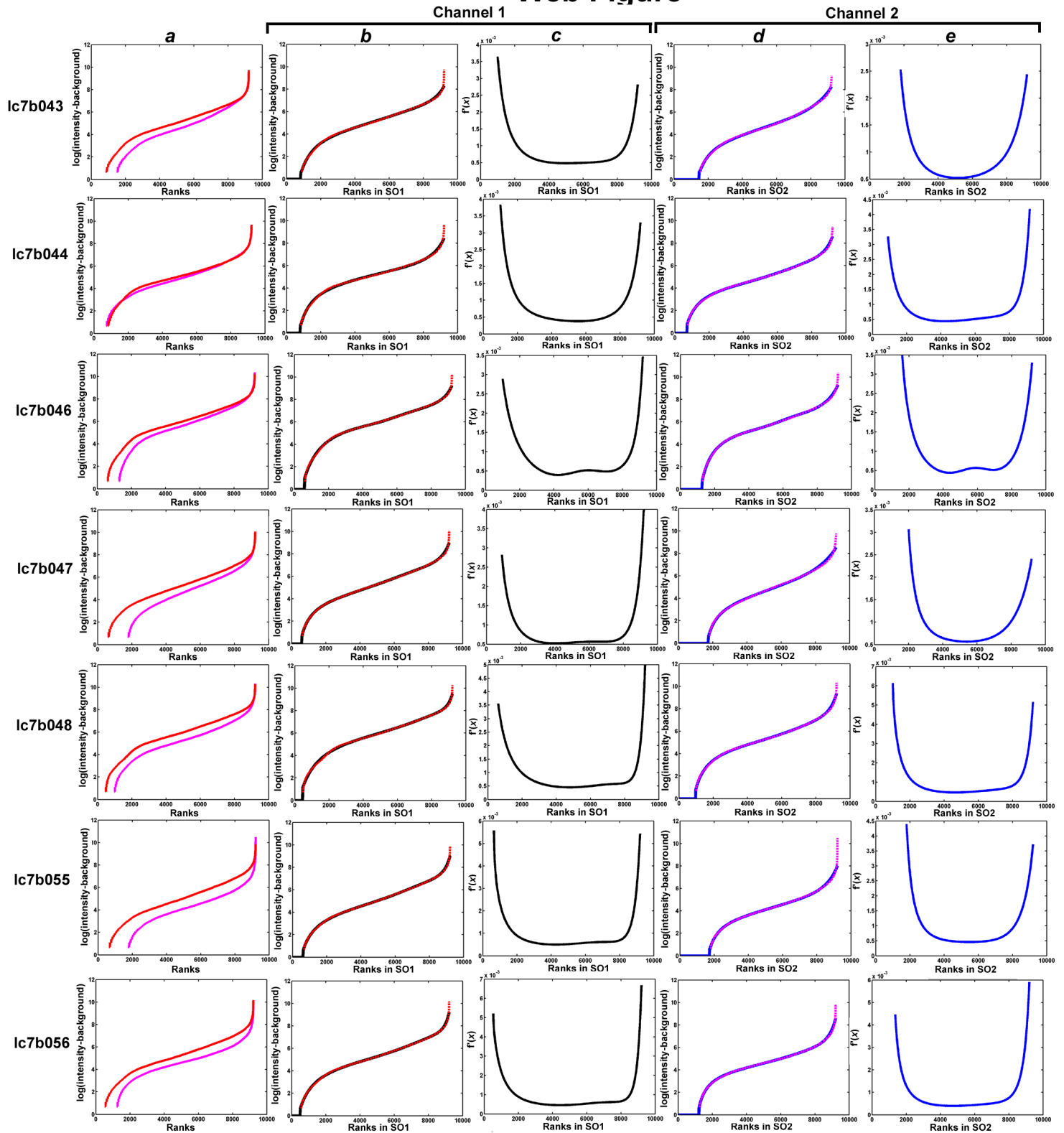
Web Figure


Channel 1
Channel 2
a
b
c
d
e
lc5b007
lc5b040
lc5b087
lc7b017
lc7b018
lc7b019
lc7b020
log(intensity-background)
f'(x)
Ranks
Ranks in SO1
Ranks in SO2

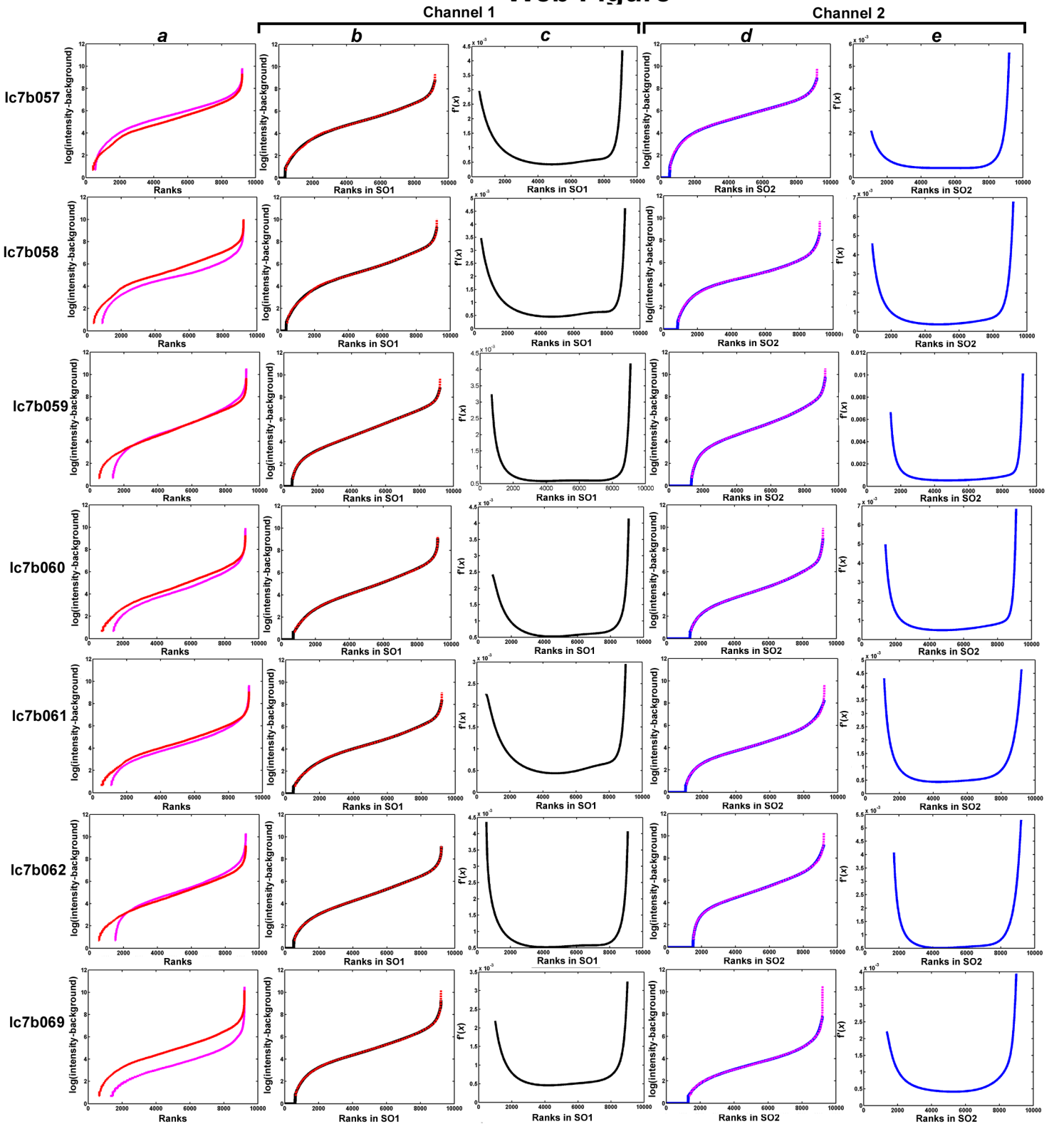
Web Figure


Channel 1
Channel 2
a
b
c
d
e
lc7b021
lc7b023
lc7b024
lc7b025
lc7b026
lc7b041
lc7b042
log(intensity-background)
f'(x)
Ranks
Ranks in SO1
Ranks in SO2

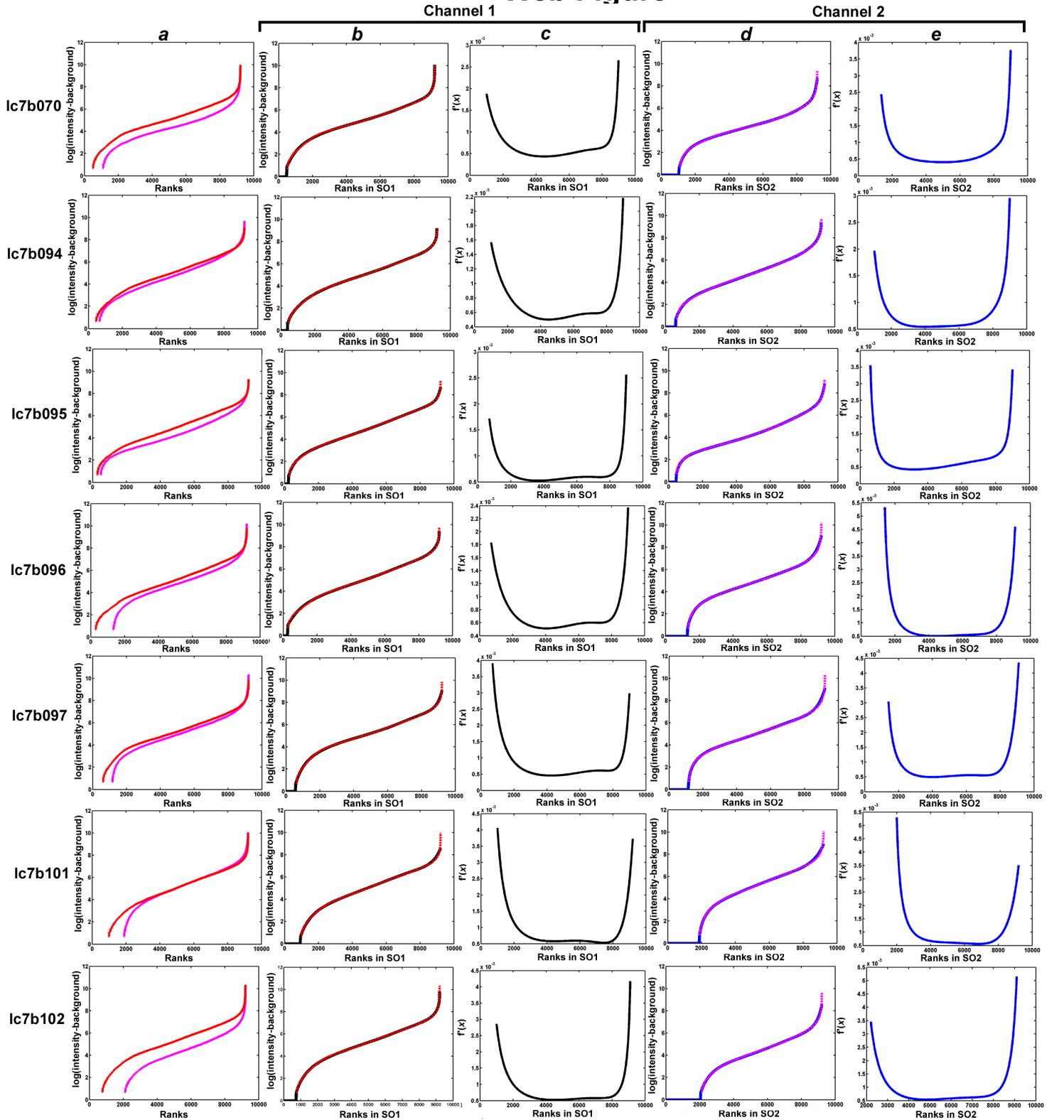
Web Figure


Channel 1
Channel 2
a
b
c
d
e
lc7b043
lc7b044
lc7b046
lc7b047
lc7b048
lc7b055
lc7b056
log(intensity-background)
f'(x)
Ranks
Ranks in SO1
Ranks in SO2

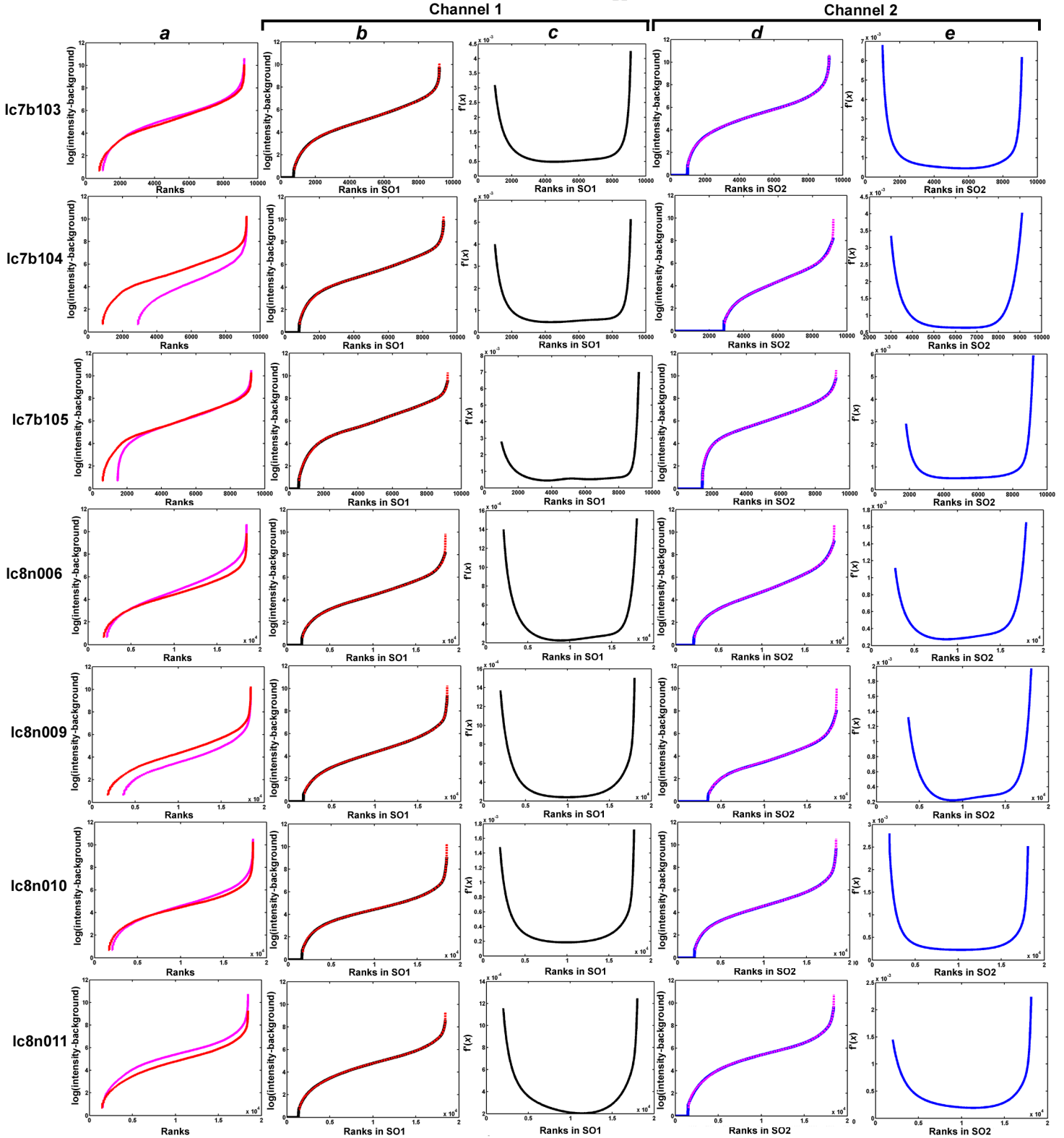
Web Figure


Channel 1
Channel 2
a
b
c
d
e
lc7b057
lc7b058
lc7b059
lc7b060
lc7b061
lc7b062
lc7b069
log(intensity-background)
f'(x)
Ranks
Ranks in SO1
Ranks in SO2

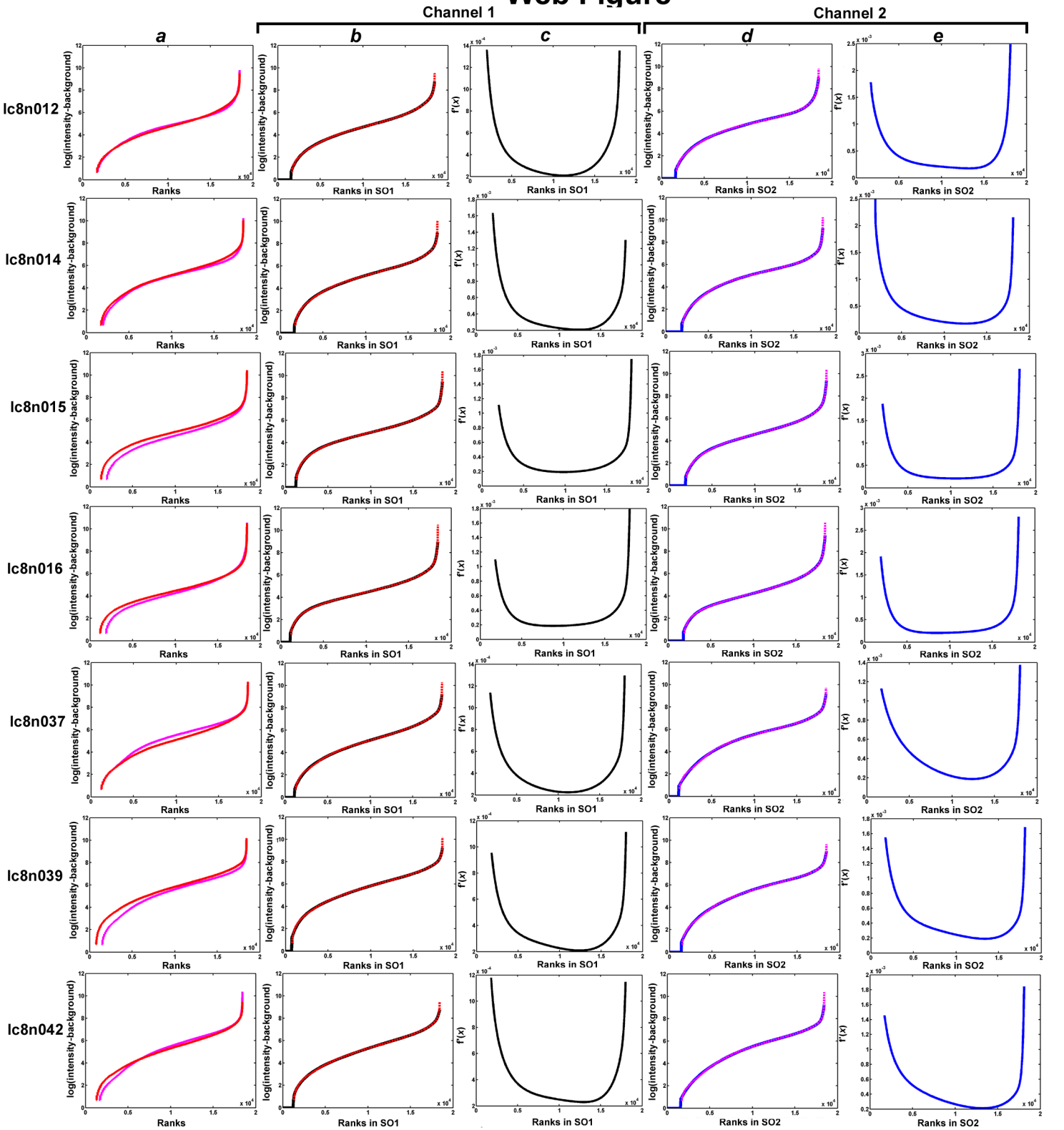
Web Figure


Channel 1
Channel 2
a
b
c
d
e
lc7b070
lc7b094
lc7b095
lc7b096
lc7b097
lc7b101
lc7b102
log(intensity-background)
f'(x)
Ranks
Ranks in SO1
Ranks in SO2

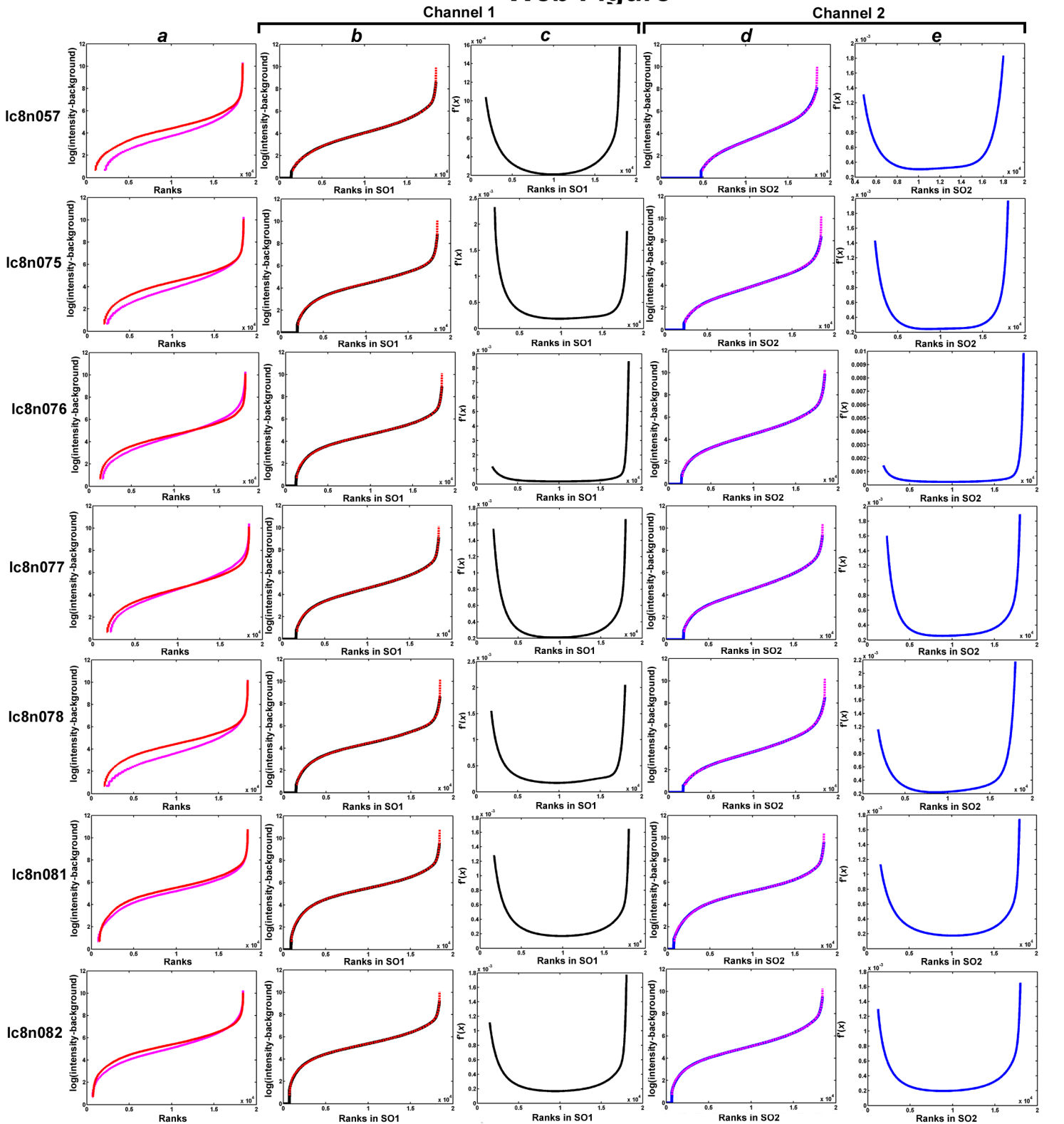
Web Figure


Channel 1
Channel 2
a
b
c
d
e
lc7b103
lc7b104
lc7b105
lc8n006
lc8n009
lc8n010
lc8n011
log(intensity-background)
f'(x)
Ranks
Ranks in SO1
Ranks in SO2

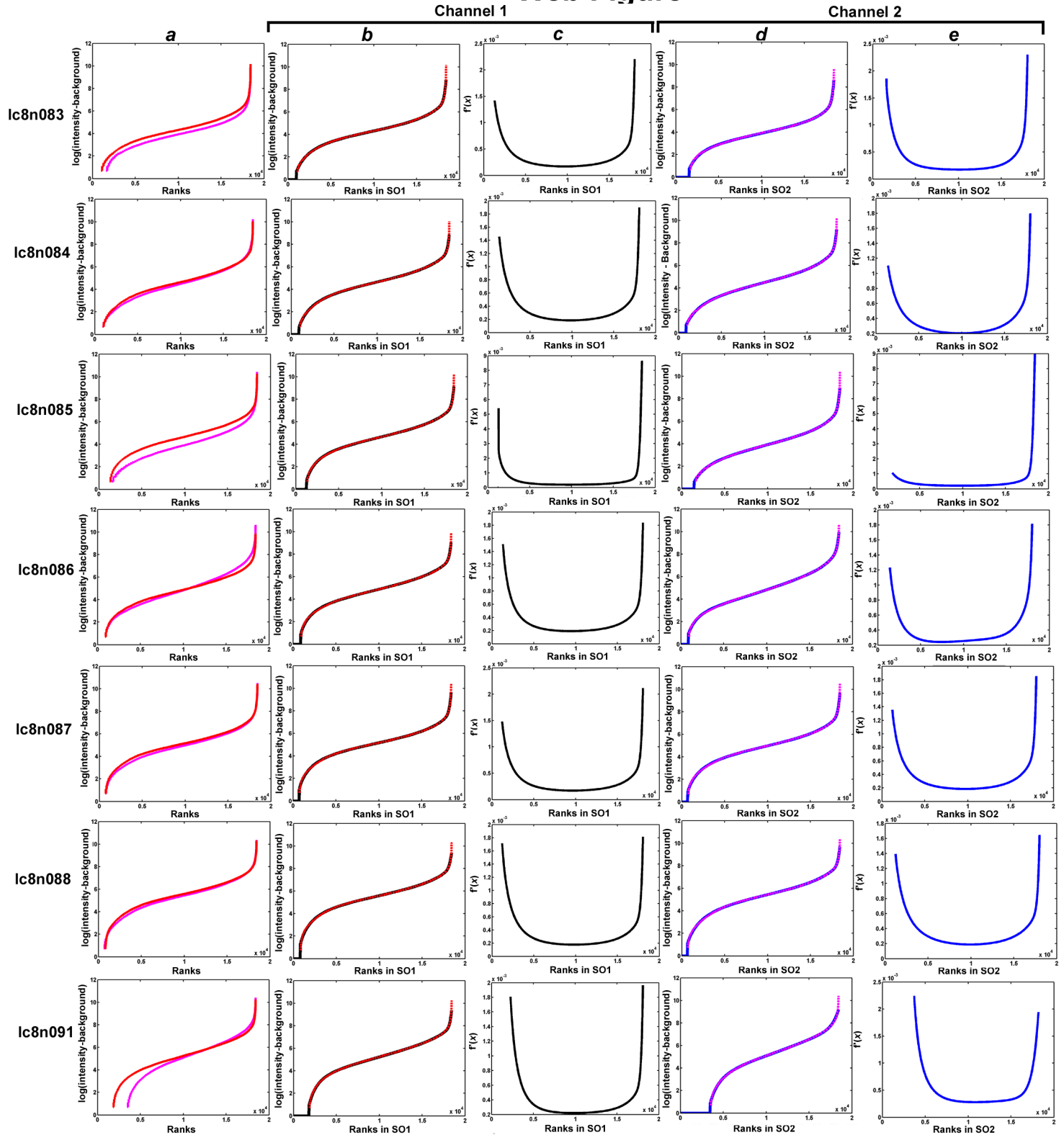
Web Figure


Channel 1
Channel 2
a
b
c
d
e
lc8n012
lc8n014
lc8n015
lc8n016
lc8n037
lc8n039
lc8n042
log(intensity-background)
f'(x)
Ranks
Ranks in SO1
Ranks in SO2

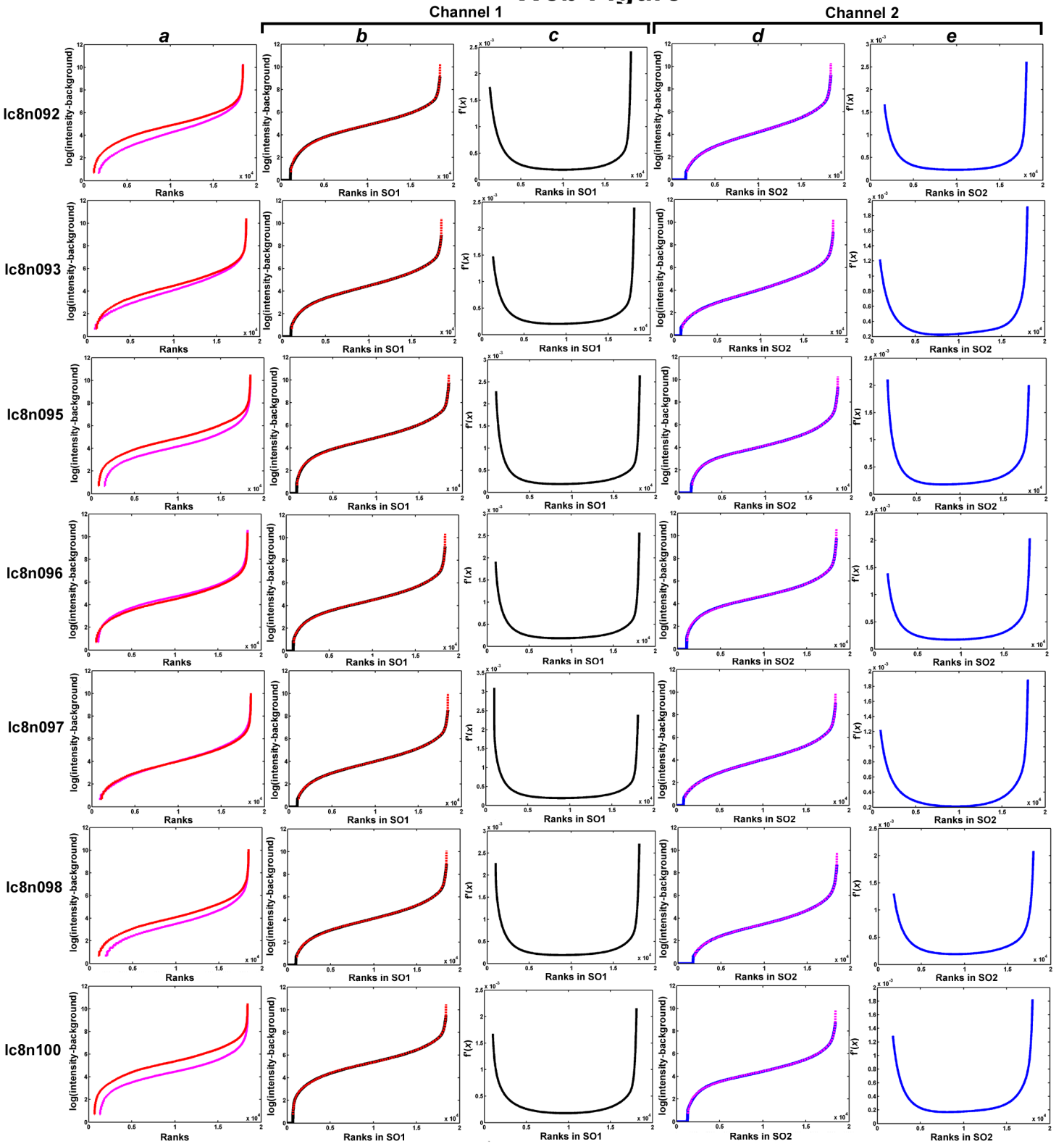
Web Figure


Channel 1
Channel 2
a
b
c
d
e
lc8n057
lc8n075
lc8n076
lc8n077
lc8n078
lc8n081
lc8n082
log(intensity-background)
f'(x)
Ranks
Ranks in SO1
Ranks in SO2

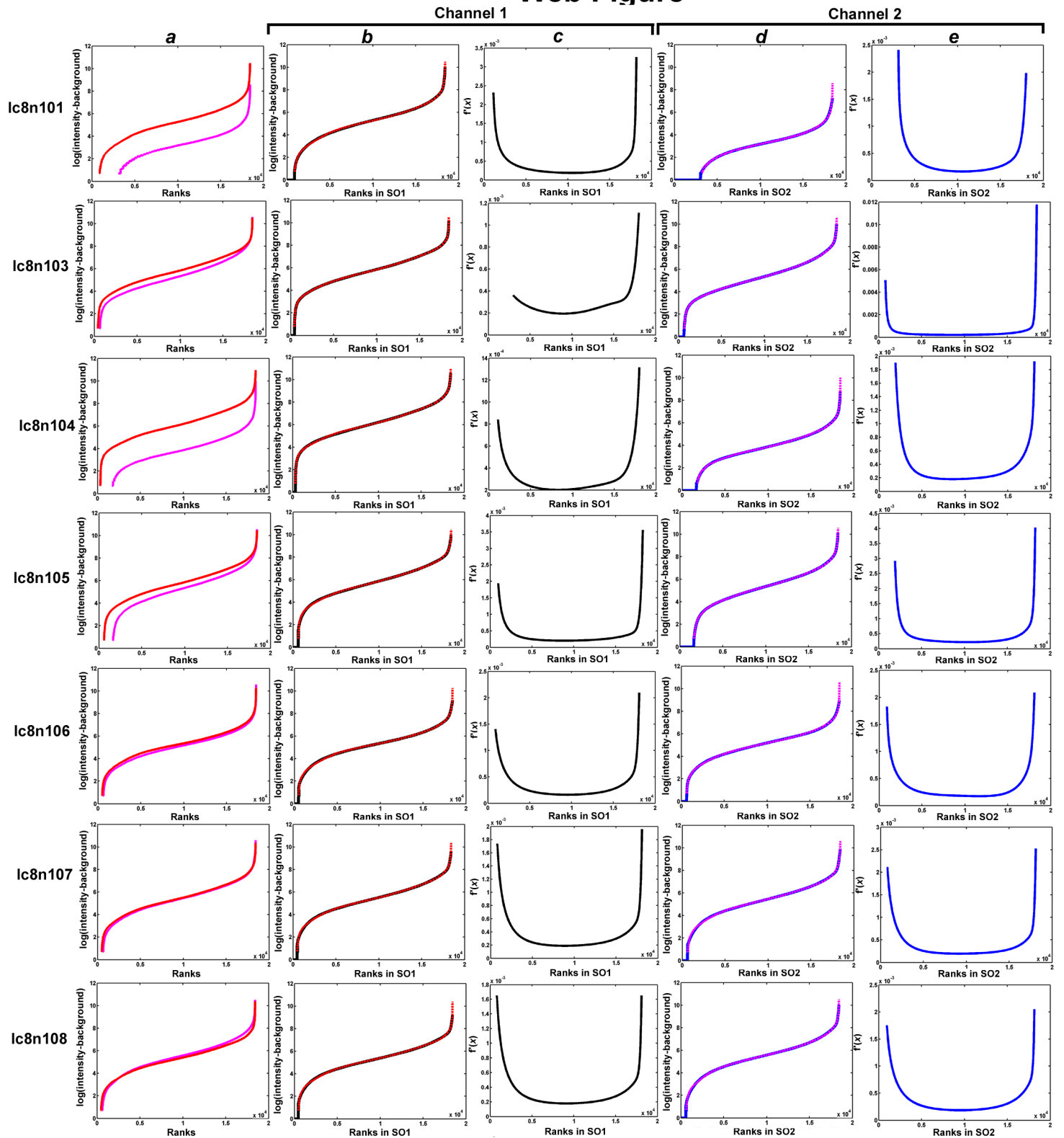
Web Figure


Channel 1
Channel 2
a
b
c
d
e
lc8n083
lc8n084
lc8n085
lc8n086
lc8n087
lc8n088
lc8n091
log(intensity-background)
log(intensity - Background)
f'(x)
Ranks
Ranks in SO1
Ranks in SO2

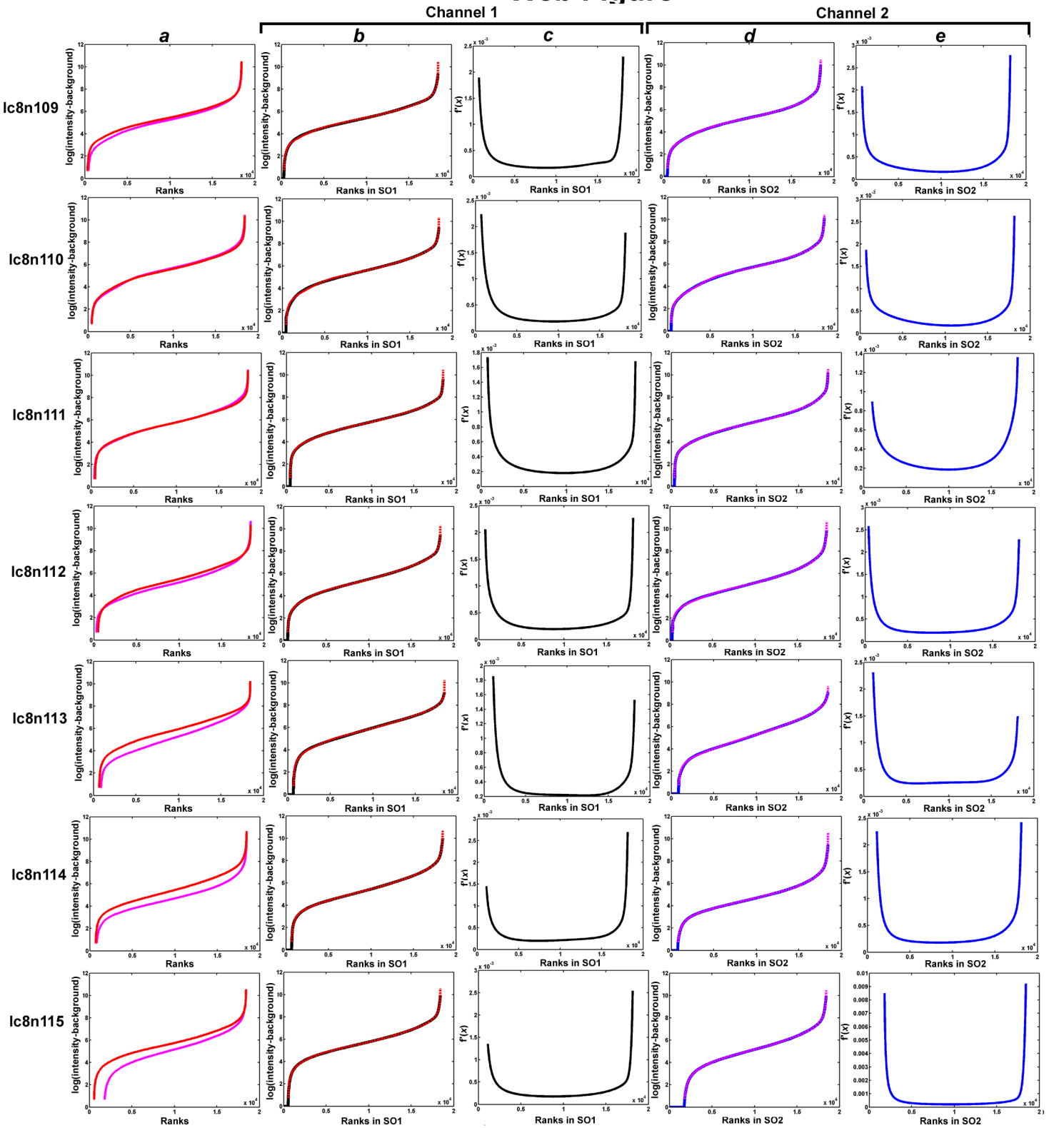
Web Figure


Channel 1
Channel 2
a
b
c
d
e
lc8n092
lc8n093
lc8n095
lc8n096
lc8n097
lc8n098
lc8n100
log(intensity-background)
f'(x)
Ranks
Ranks in SO1
Ranks in SO2

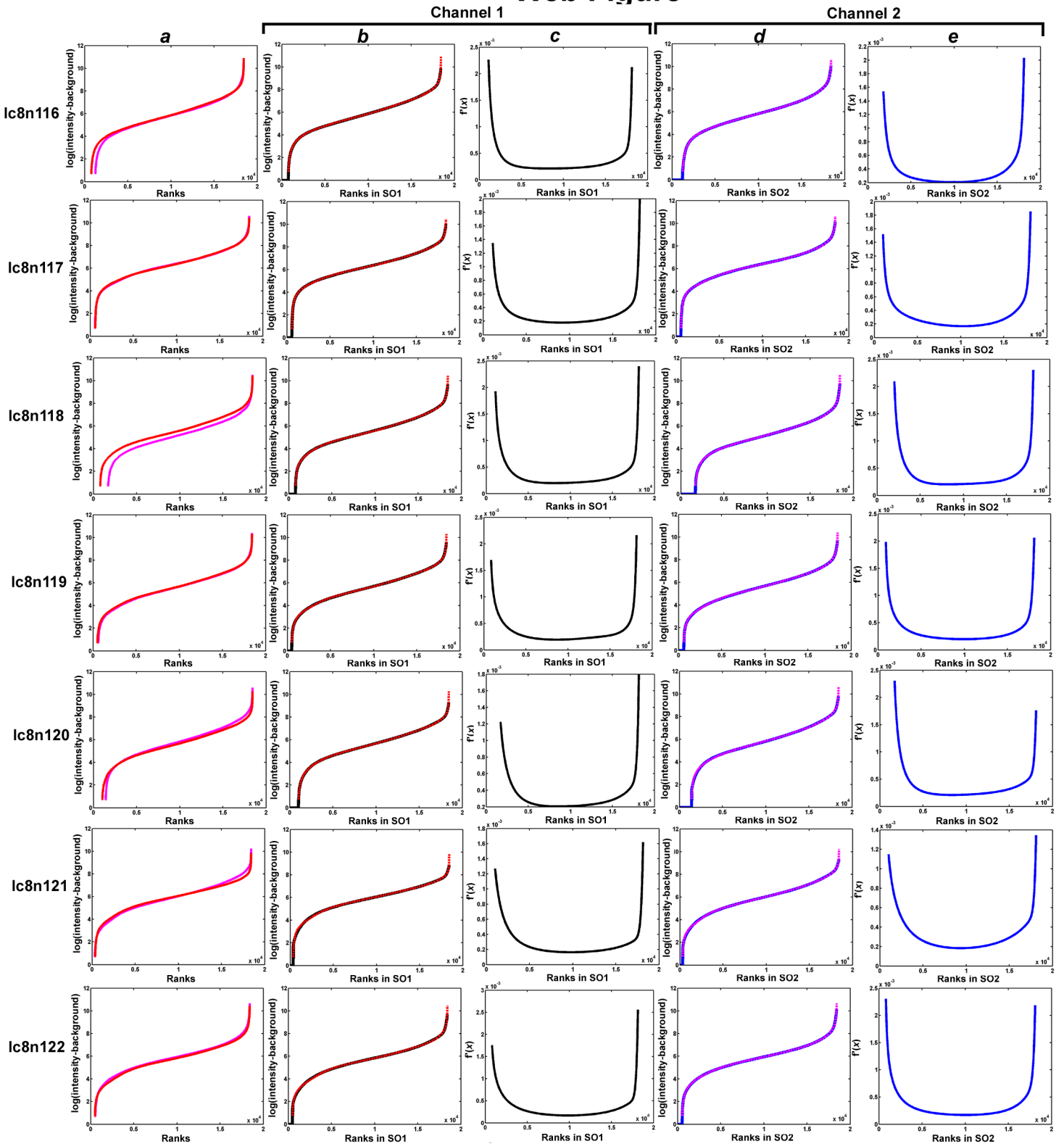
Web Figure


Channel 1
Channel 2
a
b
c
d
e
lc8n101
lc8n103
lc8n104
lc8n105
lc8n106
lc8n107
lc8n108
log(intensity-background)
f'(x)
Ranks
Ranks in SO1
Ranks in SO2

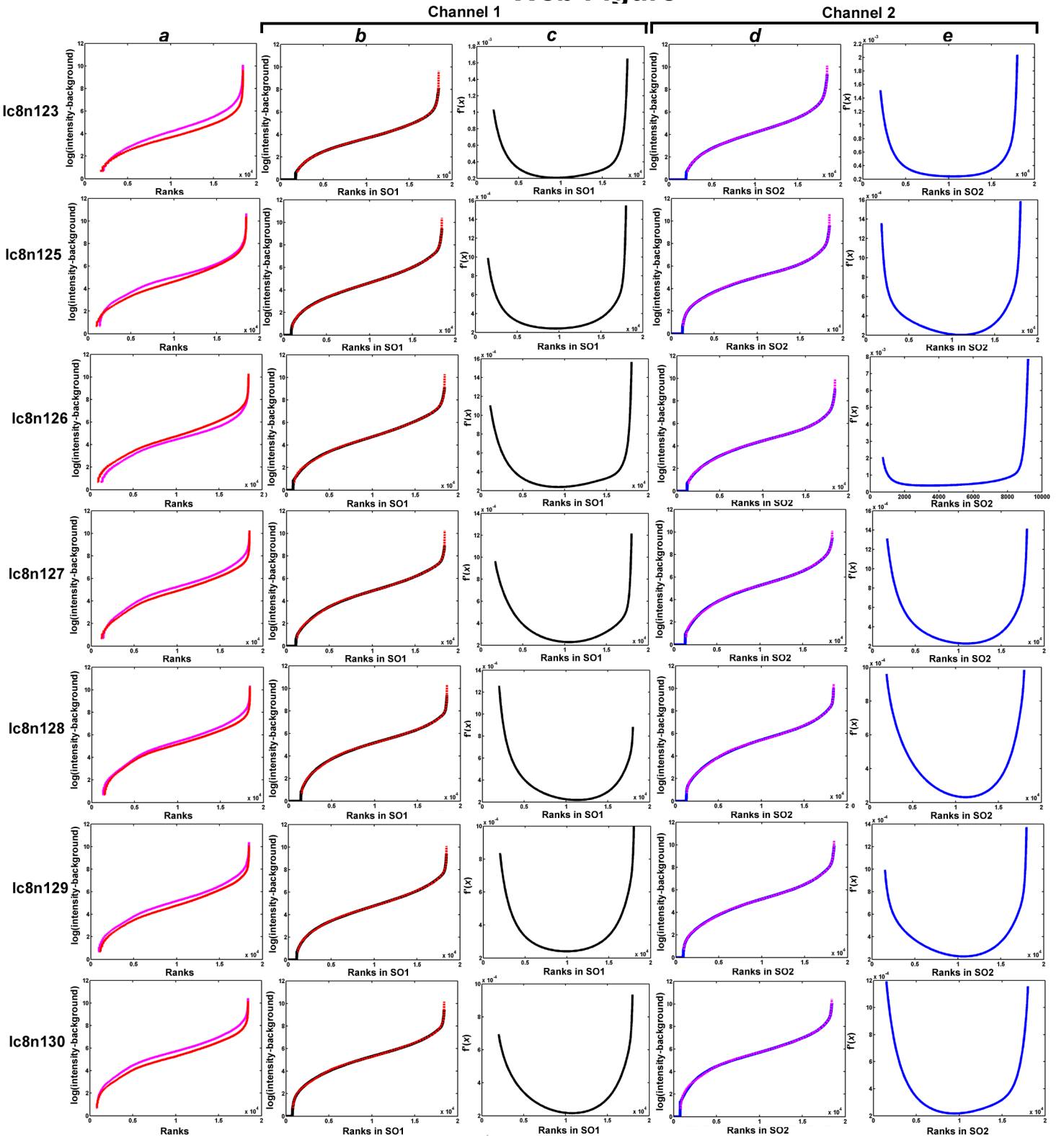
Web Figure


Channel 1
Channel 2
a
b
c
d
e
lc8n109
lc8n110
lc8n111
lc8n112
lc8n113
lc8n114
lc8n115
log(intensity-background)
f'(x)
Ranks
Ranks in SO1
Ranks in SO2

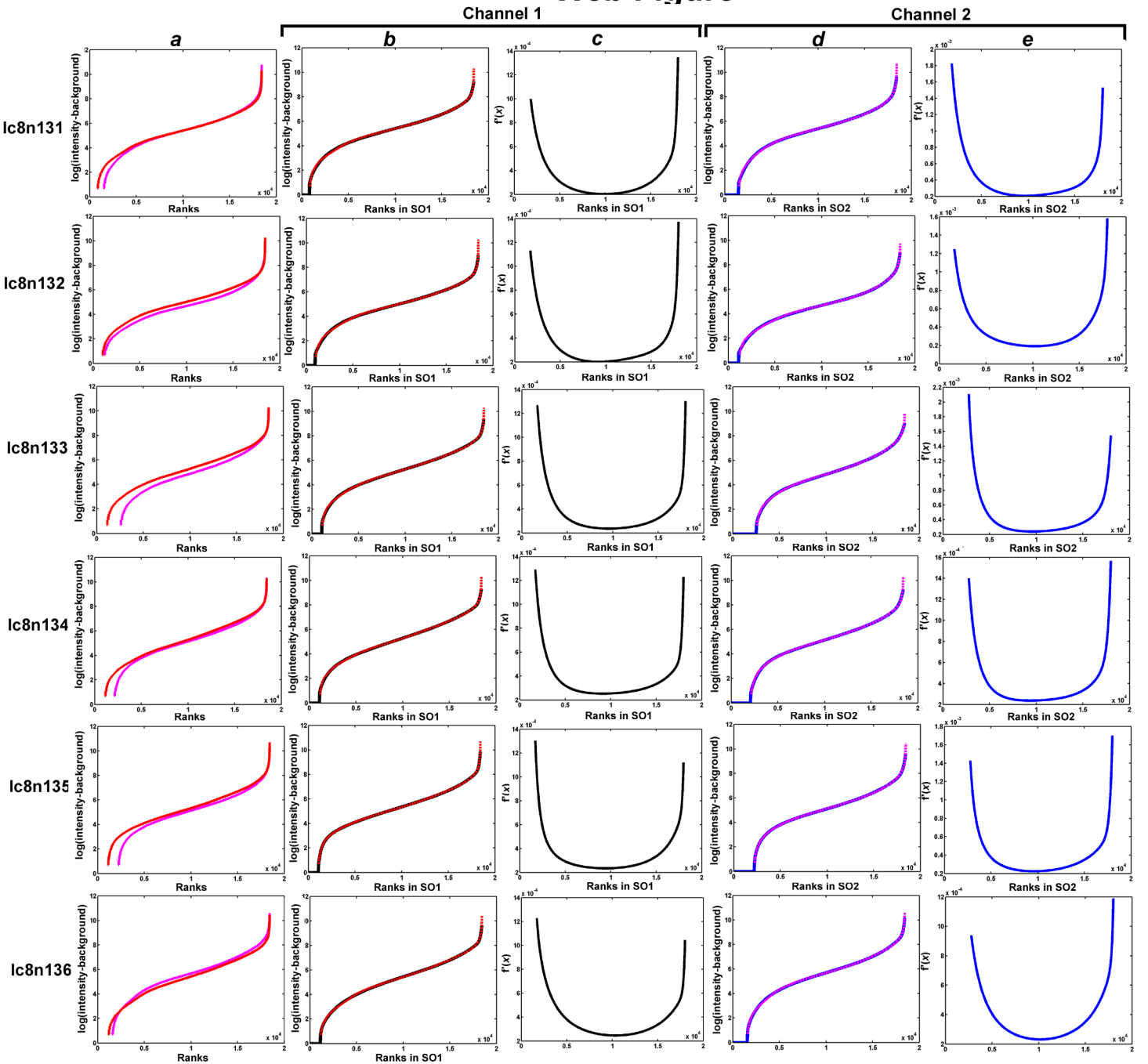
Web Figure


Channel 1
Channel 2
a
b
c
d
e
lc8n116
lc8n117
lc8n118
lc8n119
lc8n120
lc8n121
lc8n122
log(intensity-background)
f'(x)
Ranks
Ranks in SO1
Ranks in SO2

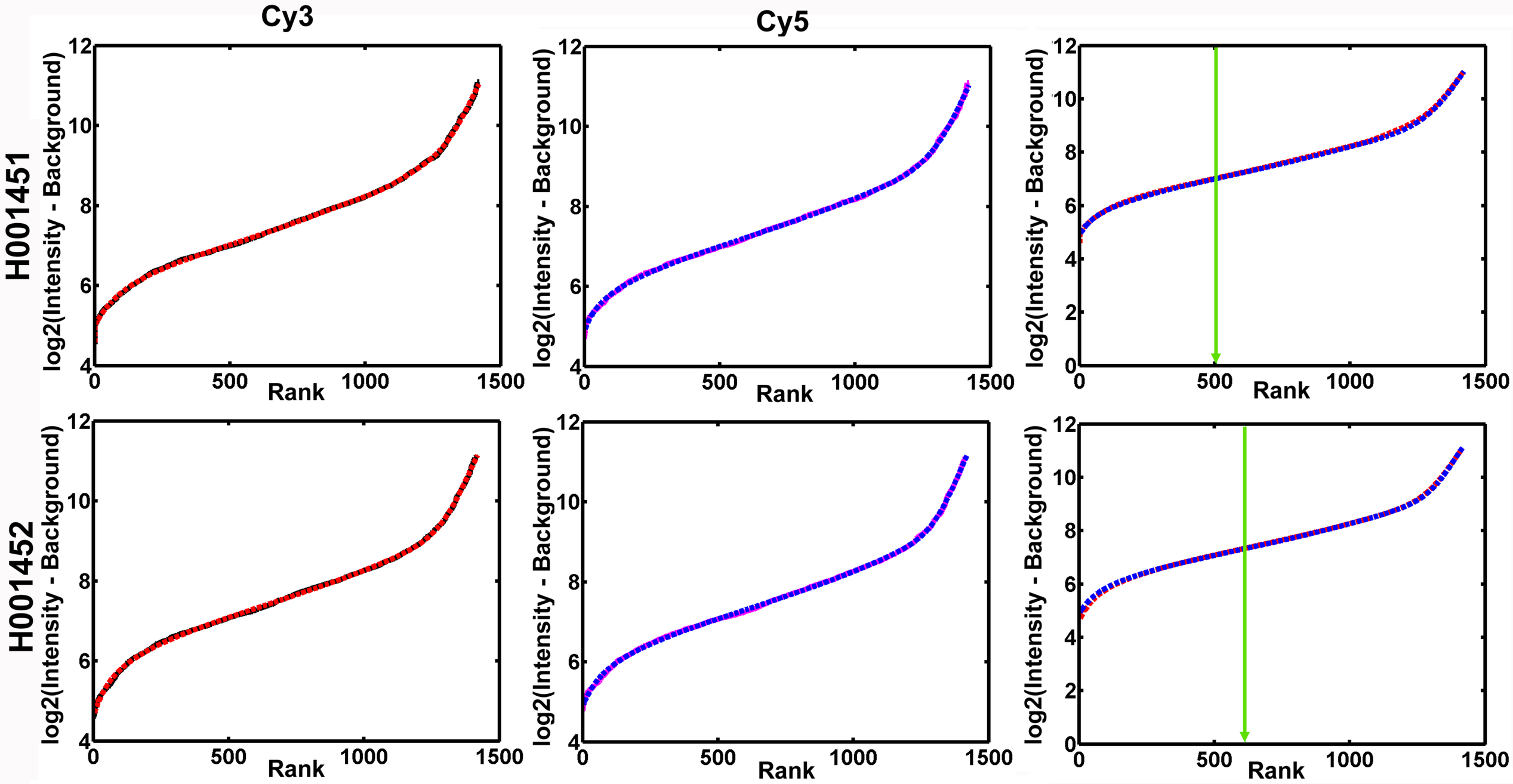
Web Figure


Channel 1
Channel 2
a
b
c
d
e
lc8n123
lc8n125
lc8n126
lc8n127
lc8n128
lc8n129
lc8n130
log(intensity-background)
f'(x)
Ranks
Ranks in SO1
Ranks in SO2

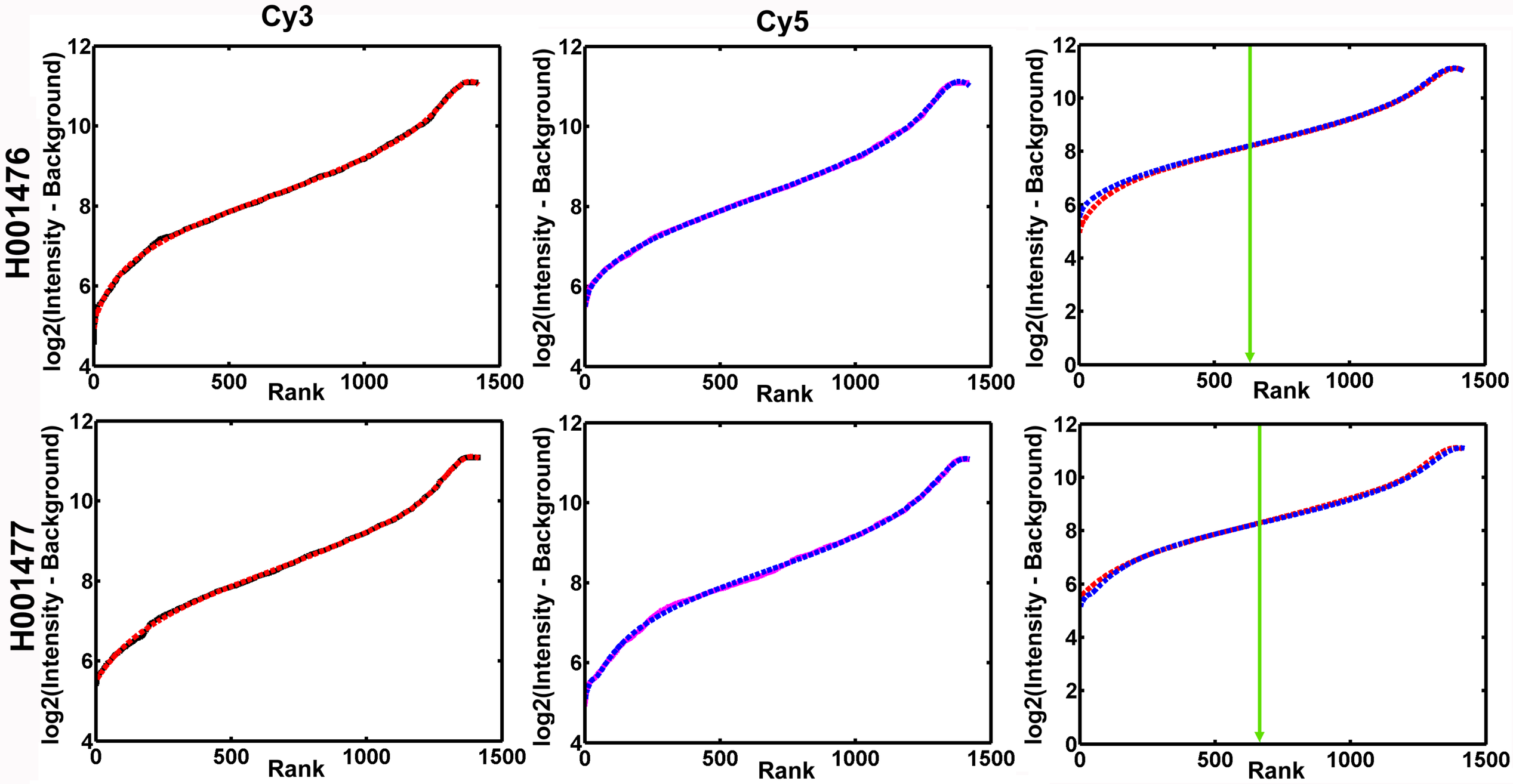
Web Figure


Channel 1
Channel 2
a
b
c
d
e
lc8n131
lc8n132
lc8n133
lc8n134
lc8n135
lc8n136
log(intensity-background)
f'(x)
Ranks
Ranks in SO1
Ranks in SO2

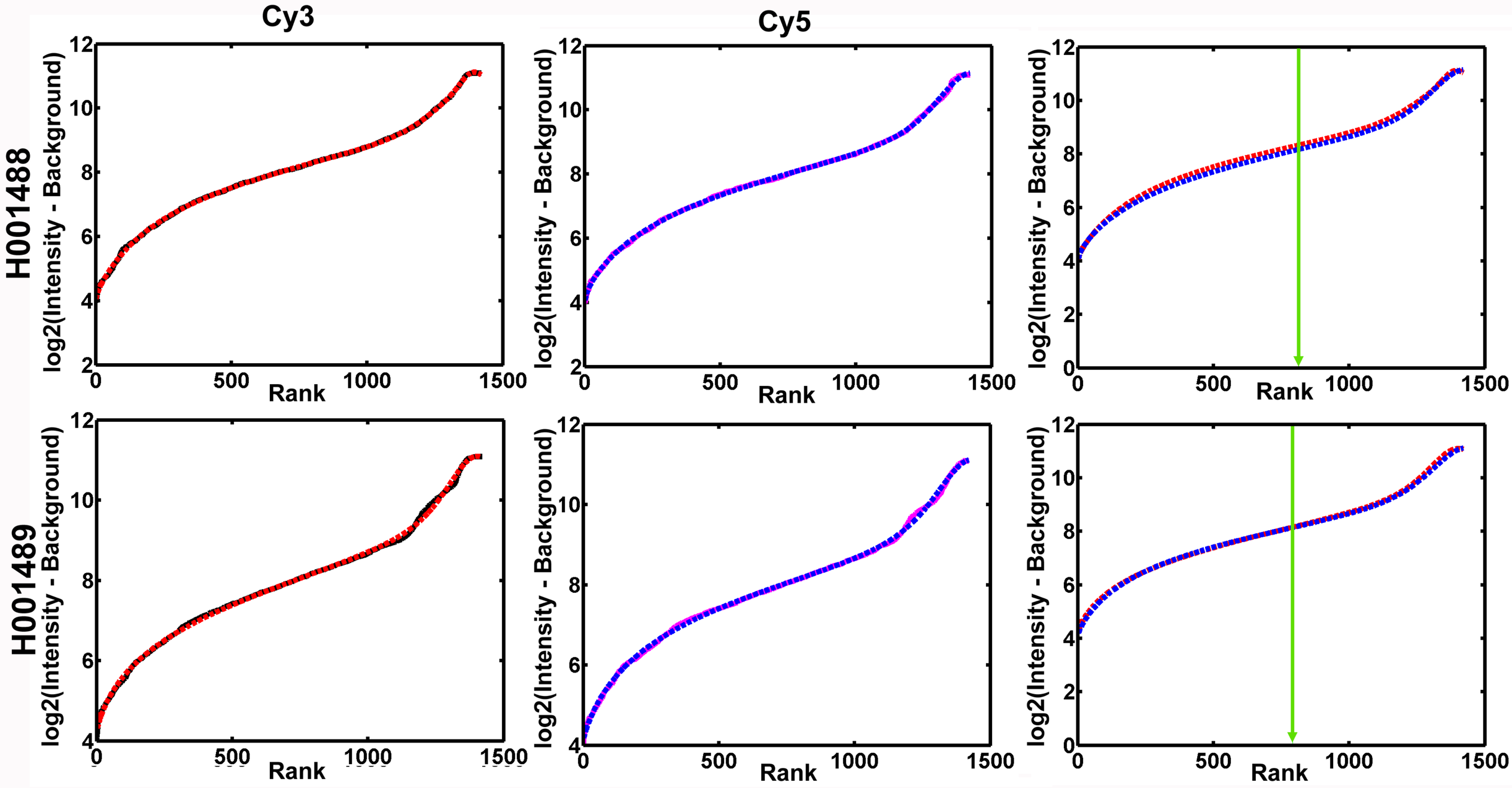
Web Figure 3


Cy3
Cy5
H001451
H001452
log2(Intensity - Background)
Rank
0
500
1000
1500
2
4
6
8
10
12

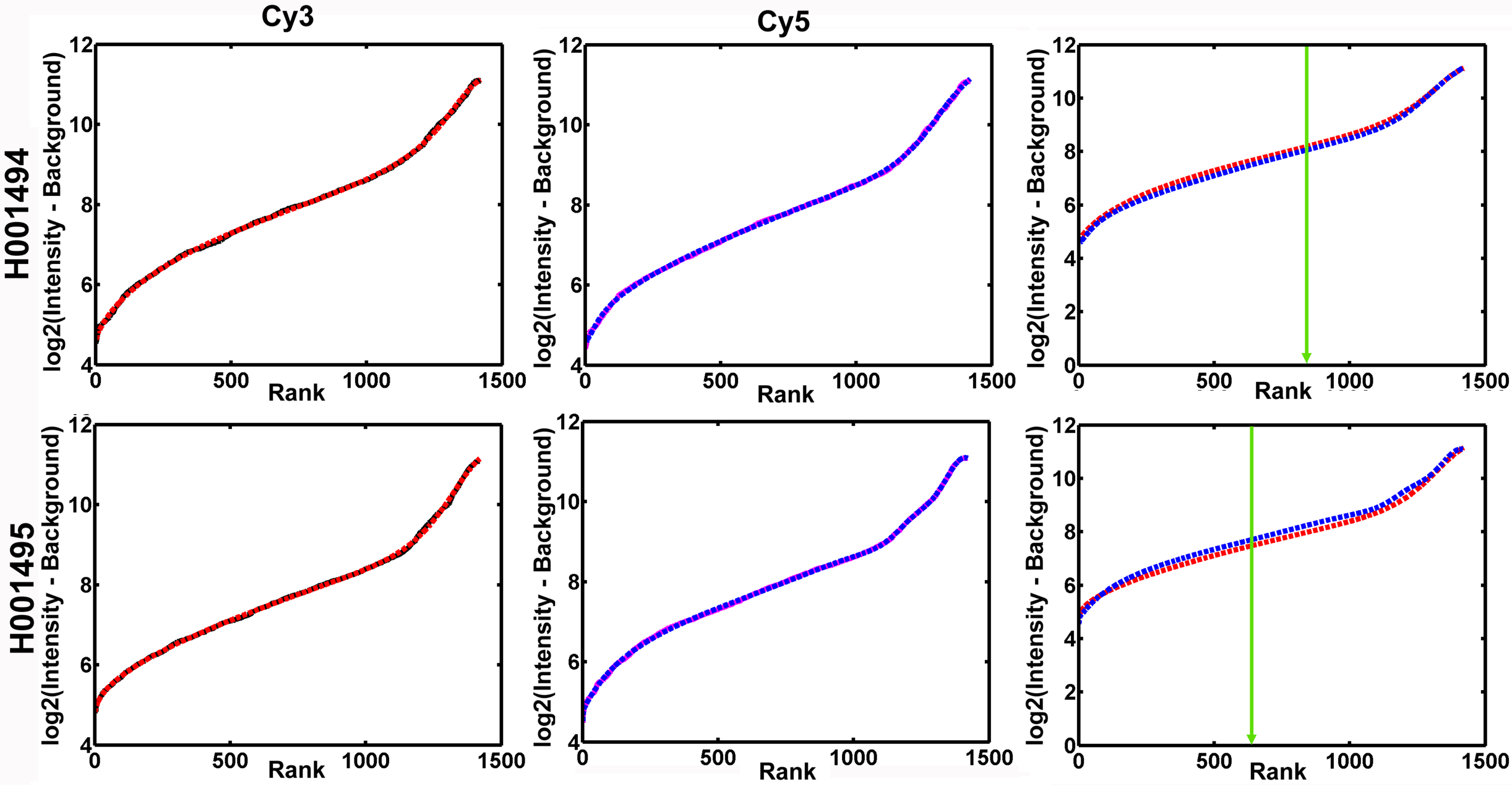
Web Figure 3


Cy3
Cy5
H001476
H001477
log2(Intensity - Background)
Rank
0
500
1000
1500
2
4
6
8
10
12

Web Figure 3


Cy3
Cy5
H001488
H001489
log2(Intensity - Background)
Rank
0
500
1000
1500
2
4
6
8
10
12

Web Figure 3


Cy3
Cy5
H001494
H001495
log2(Intensity - Background)
Rank
0
500
1000
1500
4
6
8
10
12
2