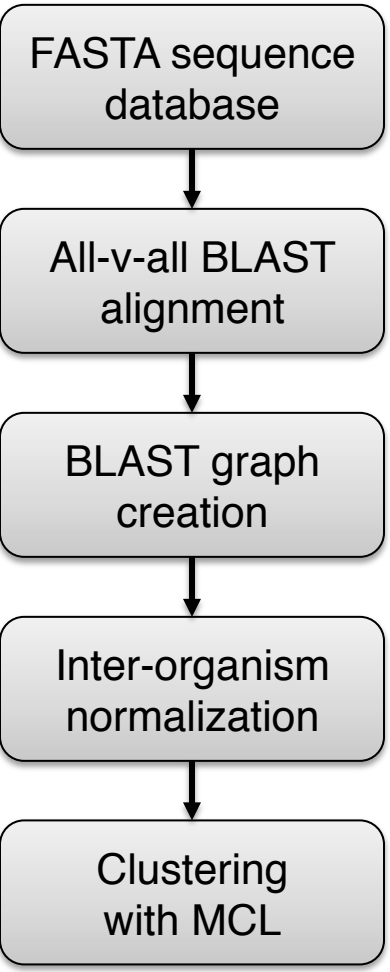
Clustering sequences using BLAST and MCL

Ted Gibbons

Overview



All sequences must be consolidated into a single FASTA file; headers must be modified to contain organism information

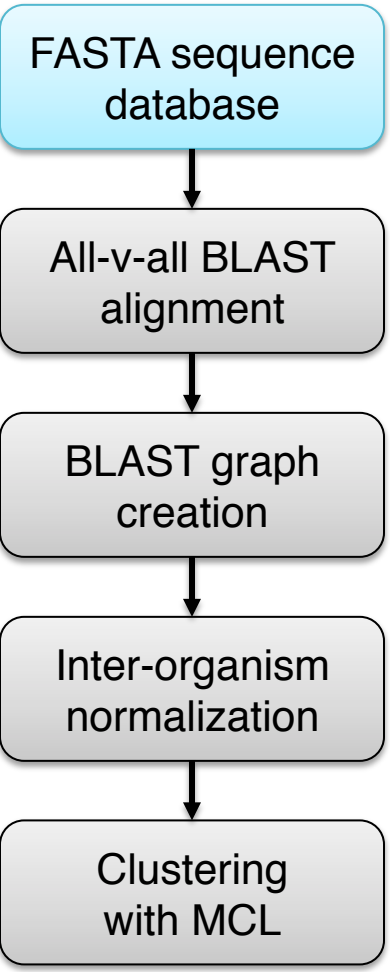
All sequences must then be aligned against each other using BLAST and printed in a machine-parsable output format

A graph is created with sequences as nodes and some BLAST-based scoring metric to weight the edges

Edges are scaled to account for differing average rates of sequence divergence between organisms

The Markov Clustering program (MCL) is then used to identify clusters of (hopefully) homologous sequences within the graph

Creating a FASTA database



harfoots.fasta

```
>seq0
WFVQMREPTLDFKASKNFMW…
>seq1
ASPTFCREYRLPGQCRKYTL…
>seq2
KWRFQNHVVHWPEVCFIWHL…
>seq3
CHYHQKVCKEKLIDRMYYDT…
>seq4
MEFTTHWQGCACN…
```

stoors.fasta

```
>seq0
SSNNATVADAWFRELSGIIG…
>seq1
QYQEWAACECMLKRRVWGDE…
>seq2
QMFPVNYNCHAQQKKKLHKV…
>seq3
VGIFPWMERDKQYLVSCVFD…
>seq4
YNWLTQEYNRVDT…
```

fallohides.fasta

```
>seq0
KENGYCMVIHQTDSFPNPLS…
>seq1
FNKPTKSPKRNCDTPMCQIA…
>seq2
SGQWDQYKKHWFVWNMWNFP…
>seq3
NMCGPHTYFRTILHPMHRMV…
>seq4
VYSWIYGIWYILY…
```

dwarves.fasta

```
>seq0
RMHVEEMACCYDGYFPGELM…
>seq1
QNWHKLEENFQRNFKRIIDL…
>seq2
CNQSTTNMIMLYKLKHCRSS…
>seq3
EALQKFWDGMTTHGICTPAQ…
>seq4
YRMRGLSFGPFAK…
```

Creating a FASTA database

FASTA sequence database

All-v-all BLAST alignment

BLAST graph creation

Inter-organism normalization

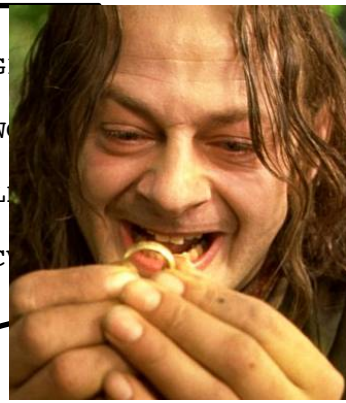
Clustering with MCL

harfoots.fasta

```
>seq0
WFVQMREPTLDFKASKNFMW…
>seq1
ASPTFCREYRLPGQCRKYTL…
VVHWPEVCFIWHL…
CKEKLIDRMYYDT…
QGCACN…
```

stoors.fasta

```
>seq0
SSNNATVADAWFRELSG
>seq1
QYQEWAACECMLKRRVW
>seq2
QMFPVNYNCHAQQKKKL
>seq3
VGIFPWMERDKQYLVSC
>seq4
YNWLTQEYNRVDT…
```



lohides.fasta

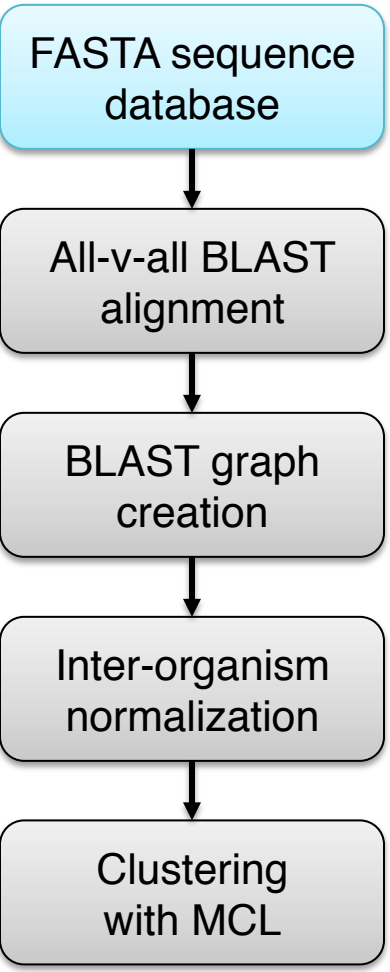
```
KENGYCMVIHQTDSFPNPLS…
>seq1
FNKPTKSPKRNCDTPMCQIA…
>seq2
SGQWDQYKKHWFVWNMWNFP…
>seq3
NMCGPHTYFRTILHPMHRMV…
>seq4
VYSWIYGIWYILY…
```

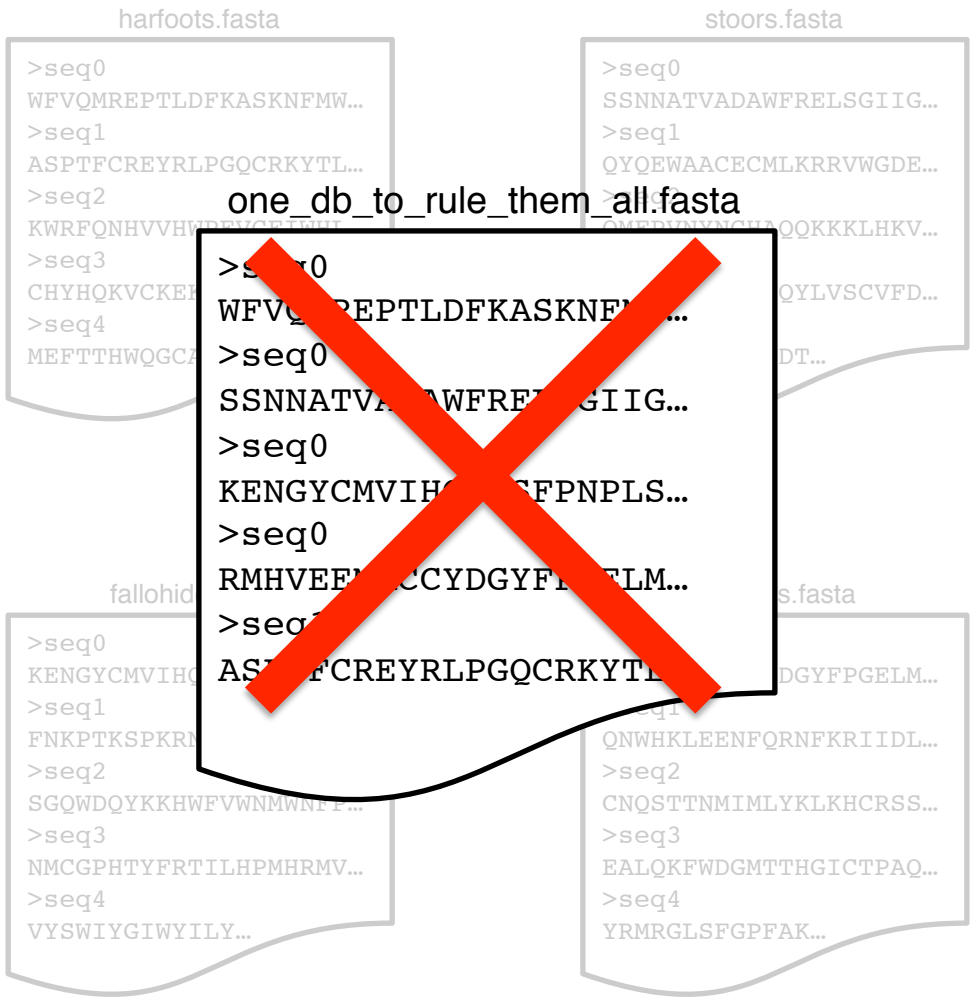
dwarves.fasta

```
>seq0
RMHVEEMACCYDGYFPG
>seq1
QNWHKLEENFQRNFKRI
>seq2
CNQSTTNMIMLYKLKHC
>seq3
EALQKFWDGMTTHGICT
>seq4
YRMRGLSFGPFAK…
```

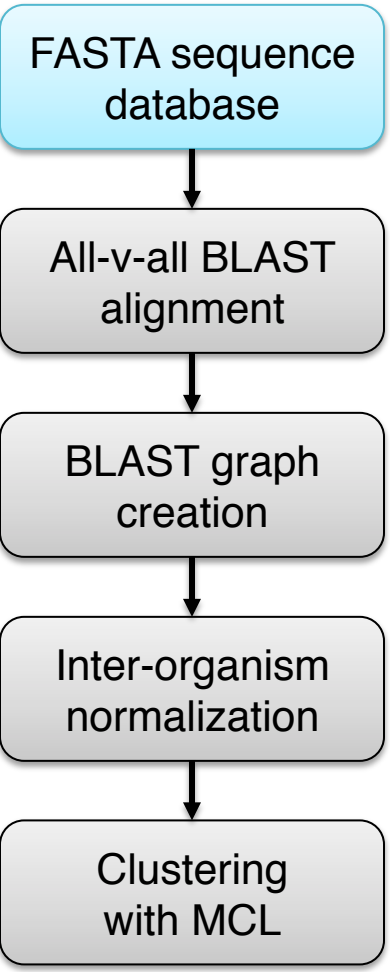
Creating a FASTA database

Creating a FASTA database





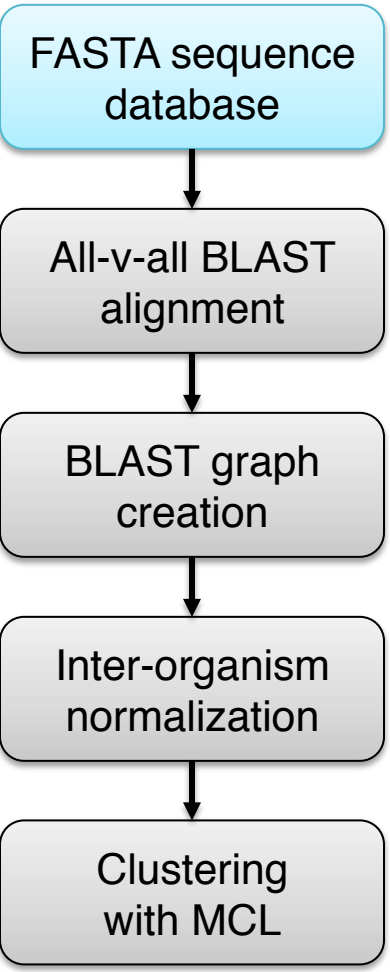
Commands

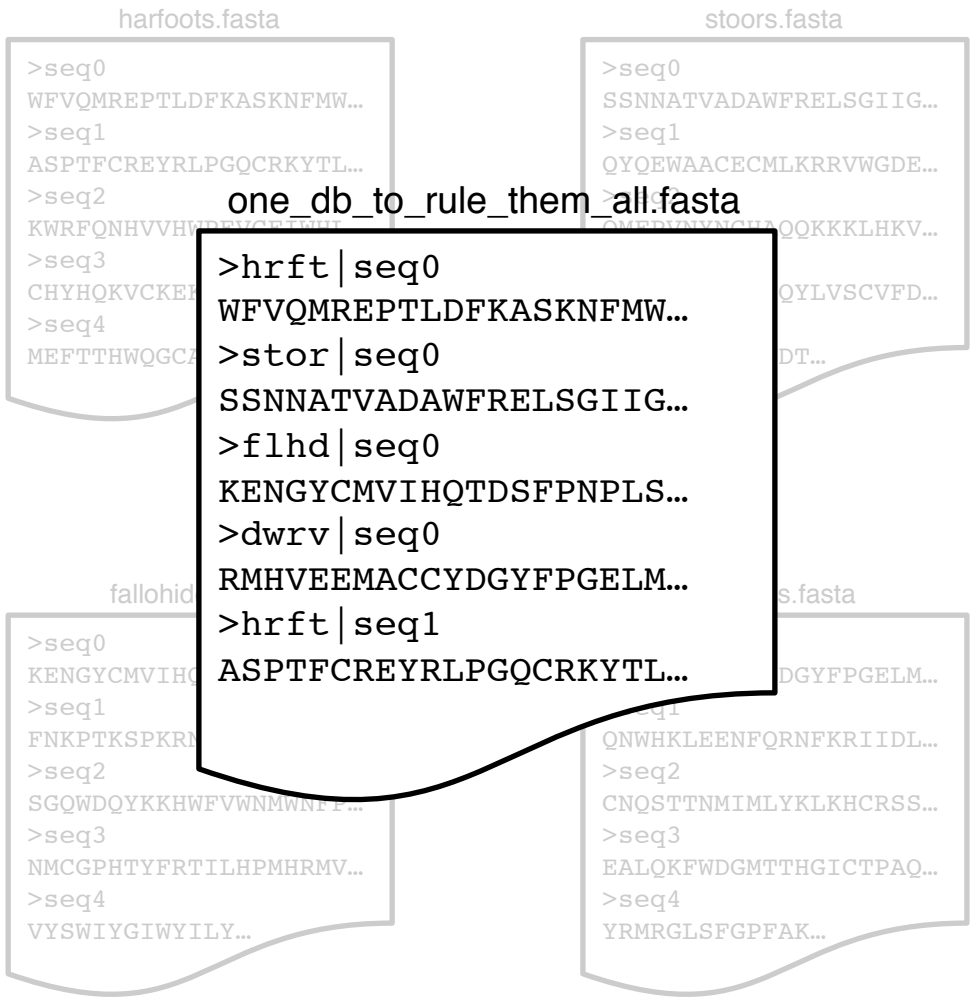



```
$ sed 's/^>/>org1ID|/g' org1.fasta > db.fasta
$ sed 's/^>/>org2ID|/g' org2.fasta >> db.fasta
$ sed 's/^>/>org3ID|/g' org3.fasta >> db.fasta
$ sed 's/^>/>org4ID|/g' org4.fasta >> db.fasta
$ sed 's/^>/>org5ID|/g' org5.fasta >> db.fasta

etc…
```

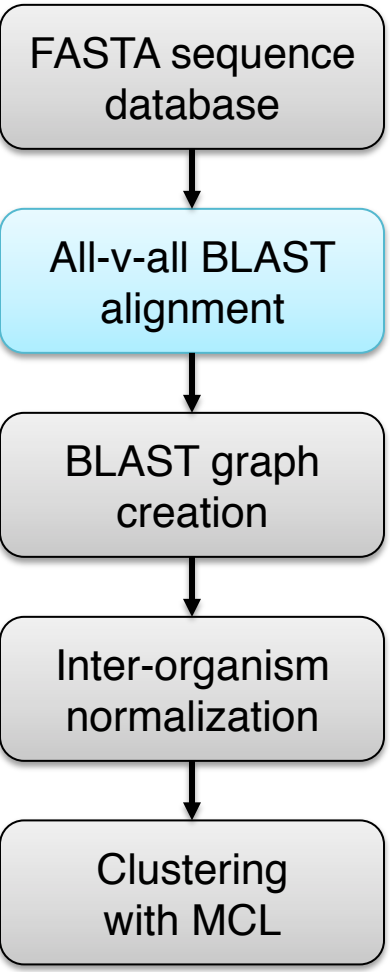
Creating a FASTA database

Creating a FASTA database



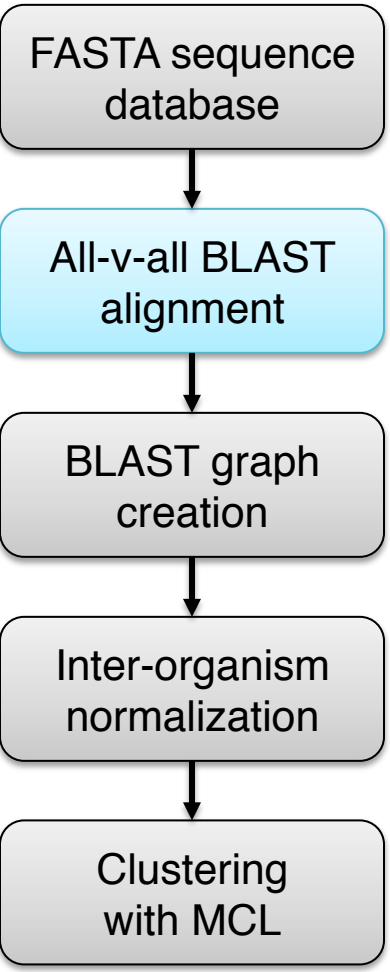


All-v-all BLAST




```
Score = 47.0 bits (110),  Expect = 1e-08,
Method: Composition-based stats., Identities = 23/44 (52%),
Positives = 33/44 (75%), Gaps = 1/44 (2%)

Query  9    IKQKLAKKLKQNRPIP-QWIRMRTGNTIRYNAKRRHWRRTKLKL  51
            ++  L ++ +++R  P Q I M+TGN IRYN+KRRHWRR+KL L
Sbjct  162  VQNHLMERPQRSRAGPNQGIWMKTGNKIRYNSKRRHWRRSKLGL  205
```

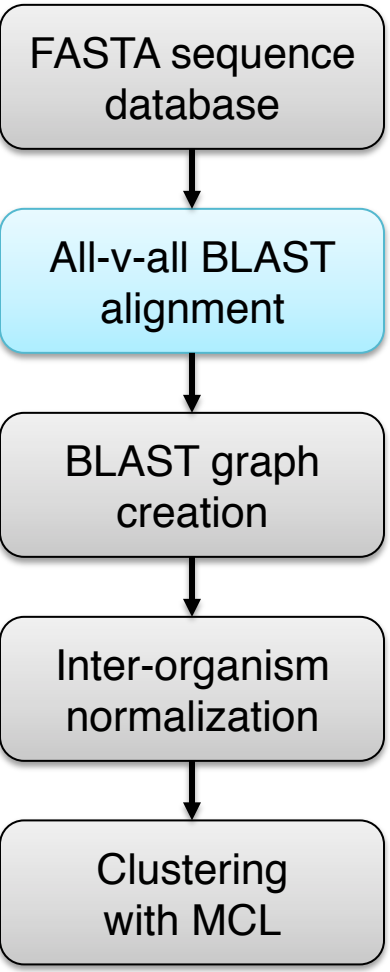
All-v-all BLAST




```
Score = 47.0 bits (110),  Expect = 1e-08,
Method: Composition-based stats., Identities = 23/44 (52%),
Positives = 33/44 (75%), Gaps = 1/44 (2%)

Query  9    IKQKLAKKLKQNKPIP-QWIRMRTGNIRYNAKRRHWRRTKLKL  51
            ++ L ++ +++R  P Q I M+TGN IRYN+KRRHWRR+KL L
Sbjct  162  VQNHLMERPQRSRAGPNQGIWMKTGNKIRYNSKRRHWRRSKLCL  205
```

```
# BLASTP 2.2.28+
# Query: hrft|seq0
# Database: one_db_to_rule_them_all.fasta
# Fields: query id, subject id, % identity, alignment length, mismatches, gap opens, q. start, q. end, s. start, ...
# 21 hits found
hrft|seq0    hrft|seq0    100.00   52       0       0       1       52       1       52       6e-31    102     52      52
hrft|seq0    stor|seq89    84.31   51       8       0       1       51       1       51       5e-25   87.8     52      51
hrft|seq0    hrft|seq3     76.47   51      12       0       1       51       1       51       1e-23   84.3     52      51
hrft|seq0    flhd|seq29    78.85   52      11       0       1       52       1       52       7e-23   82.0     52      52
hrft|seq0    hrft|seq2     76.47   51      12       0       1       51       1       51       4e-22   80.1     52      51
hrft|seq0    stor|seq15    74.51   51      13       0       1       51       1       51       2e-21   78.2     52      51
hrft|seq0    flhd|seq397   88.24   51       6       0       1       51       1       51       1e-20   76.3     52      51
hrft|seq0    hrft|seq12    69.23   52      16       0       1       52       1       52       4e-19   72.4     52      52
hrft|seq0    flhd|seq19    79.07   43       9       0       9       51       2       44       5e-19   72.0     52      44
hrft|seq0    stor|seq6     80.49   41       8       0       9       49       2       42       5e-18   69.7     52      44
hrft|seq0    hrft|seq59    74.51   51      13       0       1       51       1       51       2e-17   67.8     52      51
hrft|seq0    flhd|seq76    82.69   52       9       0       1       52       1       52       2e-16   65.5     52      52
hrft|seq0    stor|seq33    65.96   47      13       1       5       51      48       91       8e-14   59.3     52      91
hrft|seq0    stor|seq0     77.78   36       8       0       6       41     153      188       1e-12   57.4     52     202
hrft|seq0    hrft|seq456   76.32   38       7       1       6       43     380      415       8e-12   57.0     52     415
hrft|seq0    hrft|seq937   76.67   30       7       0      22       51     543      572       8e-11   54.7     52     572
hrft|seq0    stor|seq52    75.00   28       7       0      24       51     248      275       2e-09   49.7     52     275
hrft|seq0    dwrv|seq16    49.02   51      26       0       1       51       1       51       4e-09   46.2     52      52
hrft|seq0    dwrv|seq150   52.27   44      20       1       9       51     162      205       1e-08   47.0     52     205
hrft|seq0    flhd|seq47    76.92   26       6       0      26       51      94      119       3e-08   44.3     52     119
hrft|seq0    hrft|seq0     73.08   26       7       0      26       51     281      306       7e-08   45.8     52     379
```

All-v-all BLAST

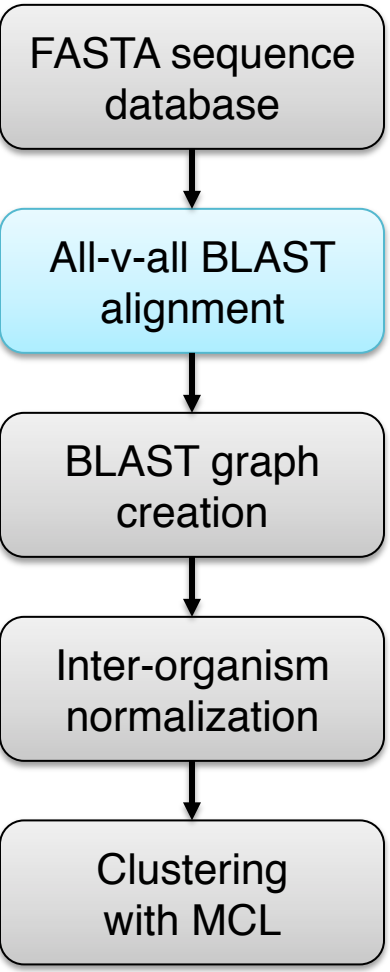



```
Score = 47.0 bits (110),  Expect = 1e-08,
Method: Composition-based stats., Identities = 23/44 (52%),
Positives = 33/44 (75%), Gaps = 1/44 (2%)

Query  9    IKQKLAKKLKQNKPIP-QWIRMRTGNIRYNAKRRHWRRTKLKL  51
            ++  L ++ +++R  P Q I M+TGN IRYN+KR HWRR+KL L
Sbjct  162  VQNHLMERPQRSRAGPNQGIWMKTGNKIRYNSKRRHWRRSKLCL  205
```

```
# BLASTP 2.2.28+
# Query: hrft|seq0
# Database: one_db_to_rule_them_all.fasta
# Fields: query id, subject id, % identity, alignment length, mismatches, gap opens, q. start, q. end, s. start, ...
# 21 hits found
```

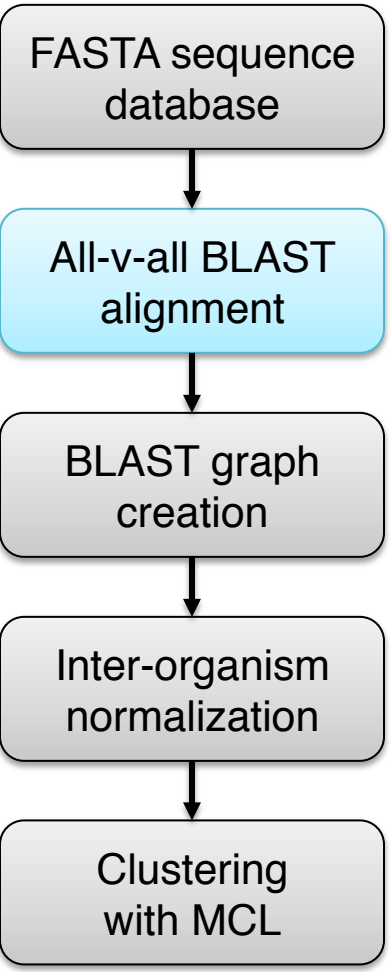
hrft|seq0	hrft|seq0	100.00	52	0	0	1	52	1	52	6e-31	102	52	52
hrft|seq0	stor|seq89	84.31	51	8	0	1	51	1	51	5e-25	87.8	52	51
hrft|seq0	hrft|seq3	76.47	51	12	0	1	51	1	51	1e-23	84.3	52	51
hrft|seq0	flhd|seq29	78.85	52	11	0	1	52	1	52	7e-23	82.0	52	52
hrft|seq0	hrft|seq2	76.47	51	12	0	1	51	1	51	4e-22	80.1	52	51
hrft|seq0	stor|seq15	74.51	51	13	0	1	51	1	51	2e-21	78.2	52	51
hrft|seq0	flhd|seq397	88.24	51	6	0	1	51	1	51	1e-20	76.3	52	51
hrft|seq0	hrft|seq12	69.23	52	16	0	1	52	1	52	4e-19	72.4	52	52
hrft|seq0	flhd|seq19	79.07	43	9	0	9	51	2	44	5e-19	72.0	52	44
hrft|seq0	stor|seq6	80.49	41	8	0	9	49	2	42	5e-18	69.7	52	44
hrft|seq0	hrft|seq59	74.51	51	13	0	1	51	1	51	2e-17	67.8	52	51
hrft|seq0	flhd|seq76	82.69	52	9	0	1	52	1	52	2e-16	65.5	52	52
hrft|seq0	stor|seq33	65.96	47	13	1	5	51	48	91	8e-14	59.3	52	91
hrft|seq0	stor|seq0	77.78	36	8	0	6	41	153	188	1e-12	57.4	52	202
hrft|seq0	hrft|seq456	76.32	38	7	1	6	43	380	415	8e-12	57.0	52	415
hrft|seq0	hrft|seq937	76.67	30	7	0	22	51	543	572	8e-11	54.7	52	572
hrft|seq0	stor|seq52	75.00	28	7	0	24	51	248	275	2e-09	49.7	52	275
hrft|seq0	dwrv|seq16	49.02	51	26	0	1	51	1	51	4e-09	46.2	52	52
hrft|seq0	dwrv|seq150	52.27	44	20	1	9	51	162	205	1e-08	47.0	52	205
hrft|seq0	flhd|seq47	76.92	26	6	0	26	51	94	119	3e-08	44.3	52	119
hrft|seq0	hrft|seq0	73.08	26	7	0	26	51	281	306	7e-08	45.8	52	379

Commands




```
$ makeblastdb -in db.fasta \
              -dbtype prot
```

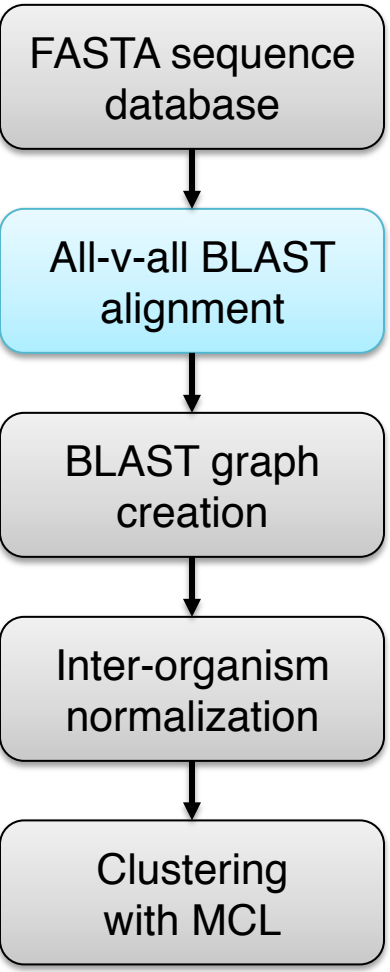
Commands




```
$ makeblastdb -in db.fasta \
              -dbtype prot
```

```
$ blastp -db test3.fasta \
         -query test3.fasta \
         -out test3.blastp \
         -evalue 1e-5 \
         -soft_masking true \
         -outfmt '7 std qlen slen' \
         -num_threads 2
```

All-v-all BLAST

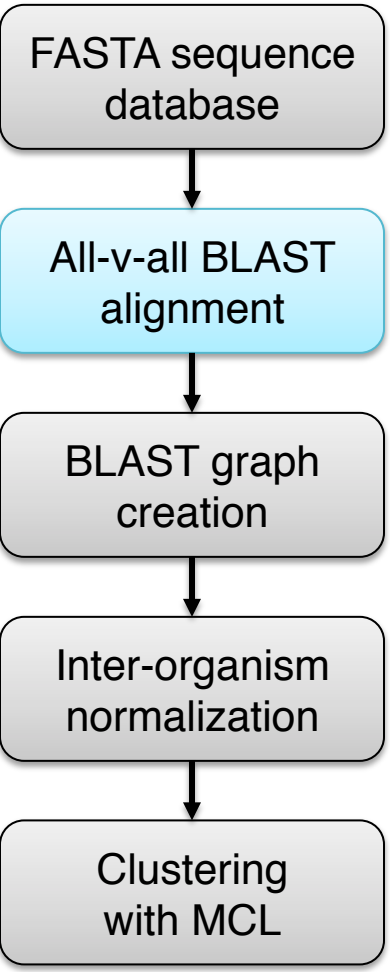



```
Score = 47.0 bits (110),  Expect = 1e-08,
Method: Composition-based stats., Identities = 23/44 (52%),
Positives = 33/44 (75%), Gaps = 1/44 (2%)

Query  9    IKQKLAKKLKQNKPIP-QWIRMRTGNIRYNAKRRHWRRTKLKL  51
            ++  L ++ +++R  P Q I M+TGN IRYN+KR HWRR+KL L
Sbjct  162  VQNHLMERPQRSRAGPNQGIWMKTGNKIRYNSKRRHWRRSKLCL  205
```

```
# BLASTP 2.2.28+
# Query: hrft|seq0
# Database: one_db_to_rule_them_all.fasta
# Fields: query id, subject id, % identity, alignment length, mismatches, gap opens, q. start, q. end, s. start, ...
# 21 hits found
```

hrft|seq0	hrft|seq0	100.00	52	0	0	1	52	1	52	6e-31	102	52	52
hrft|seq0	stor|seq89	84.31	51	8	0	1	51	1	51	5e-25	87.8	52	51
hrft|seq0	hrft|seq3	76.47	51	12	0	1	51	1	51	1e-23	84.3	52	51
hrft|seq0	flhd|seq29	78.85	52	11	0	1	52	1	52	7e-23	82.0	52	52
hrft|seq0	hrft|seq2	76.47	51	12	0	1	51	1	51	4e-22	80.1	52	51
hrft|seq0	stor|seq15	74.51	51	13	0	1	51	1	51	2e-21	78.2	52	51
hrft|seq0	flhd|seq397	88.24	51	6	0	1	51	1	51	1e-20	76.3	52	51
hrft|seq0	hrft|seq12	69.23	52	16	0	1	52	1	52	4e-19	72.4	52	52
hrft|seq0	flhd|seq19	79.07	43	9	0	9	51	2	44	5e-19	72.0	52	44
hrft|seq0	stor|seq6	80.49	41	8	0	9	49	2	42	5e-18	69.7	52	44
hrft|seq0	hrft|seq59	74.51	51	13	0	1	51	1	51	2e-17	67.8	52	51
hrft|seq0	flhd|seq76	82.69	52	9	0	1	52	1	52	2e-16	65.5	52	52
hrft|seq0	stor|seq33	65.96	47	13	1	5	51	48	91	8e-14	59.3	52	91
hrft|seq0	stor|seq0	77.78	36	8	0	6	41	153	188	1e-12	57.4	52	202
hrft|seq0	hrft|seq456	76.32	38	7	1	6	43	380	415	8e-12	57.0	52	415
hrft|seq0	hrft|seq937	76.67	30	7	0	22	51	543	572	8e-11	54.7	52	572
hrft|seq0	stor|seq52	75.00	28	7	0	24	51	248	275	2e-09	49.7	52	275
hrft|seq0	dwrv|seq16	49.02	51	26	0	1	51	1	51	4e-09	46.2	52	52
hrft|seq0	dwrv|seq150	52.27	44	20	1	9	51	162	205	1e-08	47.0	52	205
hrft|seq0	flhd|seq47	76.92	26	6	0	26	51	94	119	3e-08	44.3	52	119
hrft|seq0	hrft|seq0	73.08	26	7	0	26	51	281	306	7e-08	45.8	52	379

All-v-all BLAST

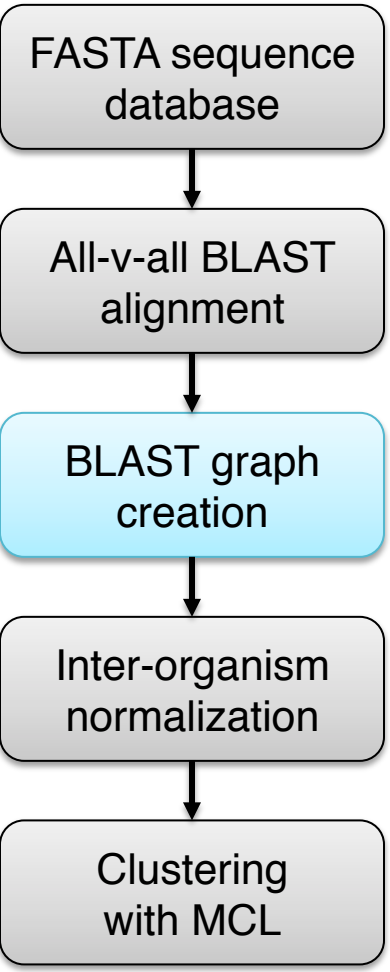



```
Score = 47.0 bits (110),  Expect = 1e-08,
Method: Composition-based stats., Identities = 23/44 (52%),
Positives = 33/44 (75%), Gaps = 1/44 (2%)

Query  9    IKQKLAKKLKQNKPIP-QWIRMRTGNIRYNAKRRHWRRTKLKL  51
            ++ L ++ +++R  P Q I M+TGN IRYN+KR HWRR+KL L
Sbjct  162  VQNHLMERPQRSRAGPNQGIWMKTGNKIRYNSKRRHWRRSKLCL  205
```

```
# BLASTP 2.2.28+
# Query: hrft|seq0
# Database: one_db_to_rule_them_all.fasta
# Fields: query id, subject id, % identity, alignment length, mismatches, gap opens, q. start, q. end, s. start, ...
# 21 hits found
```

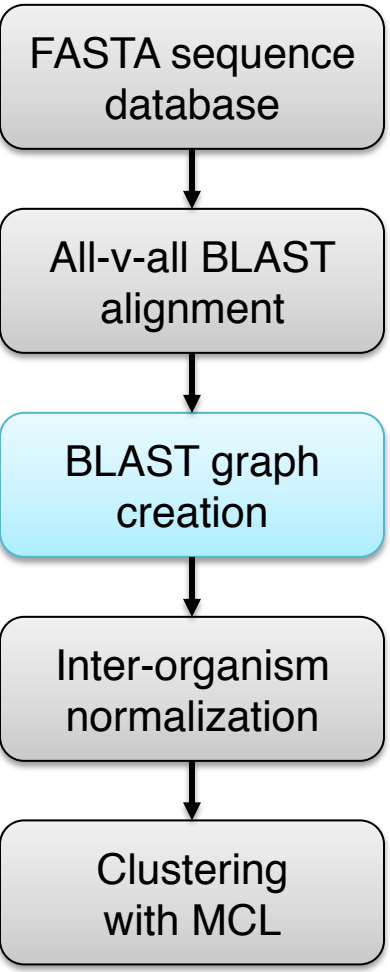
hrft|seq0	hrft|seq0	100.00	52	0	0	1	52	1	52	6e-31	102	52	52
hrft|seq0	stor|seq89	84.31	51	8	0	1	51	1	51	5e-25	87.8	52	51
hrft|seq0	hrft|seq3	76.47	51	12	0	1	51	1	51	1e-23	84.3	52	51
hrft|seq0	flhd|seq29	78.85	52	11	0	1	52	1	52	7e-23	82.0	52	52
hrft|seq0	hrft|seq2	76.47	51	12	0	1	51	1	51	4e-22	80.1	52	51
hrft|seq0	stor|seq15	74.51	51	13	0	1	51	1	51	2e-21	78.2	52	51
hrft|seq0	flhd|seq397	88.24	51	6	0	1	51	1	51	1e-20	76.3	52	51
hrft|seq0	hrft|seq12	69.23	52	16	0	1	52	1	52	4e-19	72.4	52	52
hrft|seq0	flhd|seq19	79.07	43	9	0	9	51	2	44	5e-19	72.0	52	44
hrft|seq0	stor|seq6	80.49	41	8	0	9	49	2	42	5e-18	69.7	52	44
hrft|seq0	hrft|seq59	74.51	51	13	0	1	51	1	51	2e-17	67.8	52	51
hrft|seq0	flhd|seq76	82.69	52	9	0	1	52	1	52	2e-16	65.5	52	52
hrft|seq0	stor|seq33	65.96	47	13	1	5	51	48	91	8e-14	59.3	52	91
hrft|seq0	stor|seq0	77.78	36	8	0	6	41	153	188	1e-12	57.4	52	202
hrft|seq0	hrft|seq456	76.32	38	7	1	6	43	380	415	8e-12	57.0	52	415
hrft|seq0	hrft|seq937	76.67	30	7	0	22	51	543	572	8e-11	54.7	52	572
hrft|seq0	stor|seq52	75.00	28	7	0	24	51	248	275	2e-09	49.7	52	275
hrft|seq0	dwrv|seq16	49.02	51	26	0	1	51	1	51	4e-09	46.2	52	52
hrft|seq0	dwrv|seq150	52.27	44	20	1	9	51	162	205	1e-08	47.0	52	205
hrft|seq0	flhd|seq47	76.92	26	6	0	26	51	94	119	3e-08	44.3	52	119
hrft|seq0	hrft|seq0	73.08	26	7	0	26	51	281	306	7e-08	45.8	52	379

BLAST graph



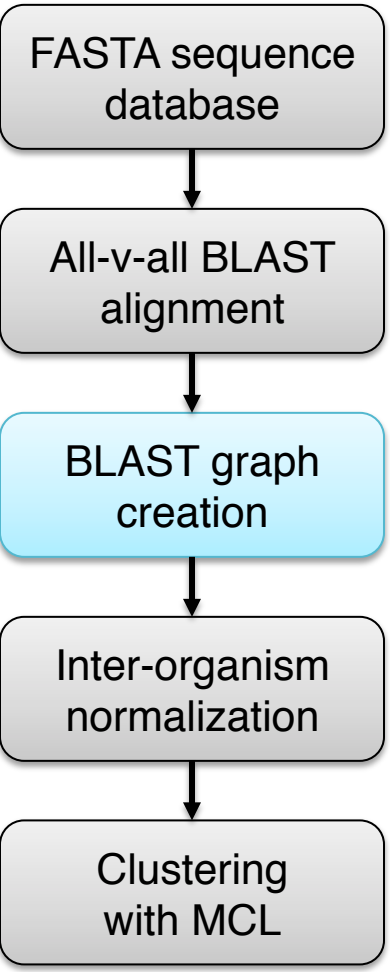
	hrft|seq0	stor|seq0	flhd|seq0	dwrv|seq0	hrft|seq1	stor|seq1	flhd|seq1
hrft|seq0	251	0	131	79	0	140	0
stor|seq0	0	165	0	0	0	96	137
flhd|seq0	131	0	282	0	119	95	147
dwrv|seq0	79	0	0	181	85	131	0
hrft|seq1	0	0	119	85	205	145	135
stor|seq1	140	96	95	131	145	253	0
flhd|seq1	0	137	147	0	135	0	238

BLAST graph



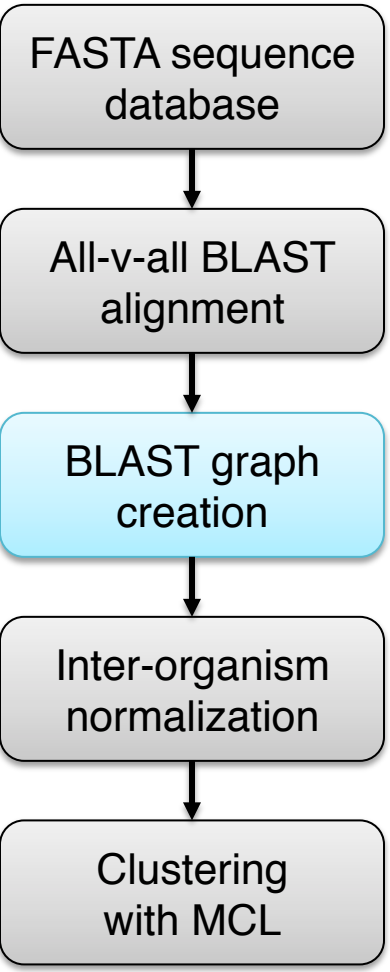
	hrft|seq0	stor|seq0	flhd|seq0	dwrv|seq0	hrft|seq1	stor|seq1	flhd|seq1
hrft|seq0	251	0	131	79	0	140	0
stor|seq0	0	165	0	0	0	96	137
flhd|seq0	131	0	282	0	119	95	147
dwrv|seq0	79	0	0	181	85	131	0
hrft|seq1	0	0	119	85	205	145	135
stor|seq1	140	96	95	131	145	253	0
flhd|seq1	0	137	147	0	135	0	238

BLAST graph

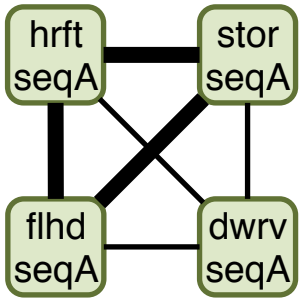


	hrft|seq0	stor|seq0	flhd|seq0	dwrv|seq0	hrft|seq1	stor|seq1	flhd|seq1
hrft|seq0	251	0	131	79	0	140	0
stor|seq0	0	165	0	0	0	96	137
flhd|seq0	131	0	282	0	119	95	147
dwrv|seq0	79	0	0	181	85	131	0
hrft|seq1	0	0	119	85	205	145	135
stor|seq1	140	96	95	131	145	253	0
flhd|seq1	0	137	147	0	135	0	238

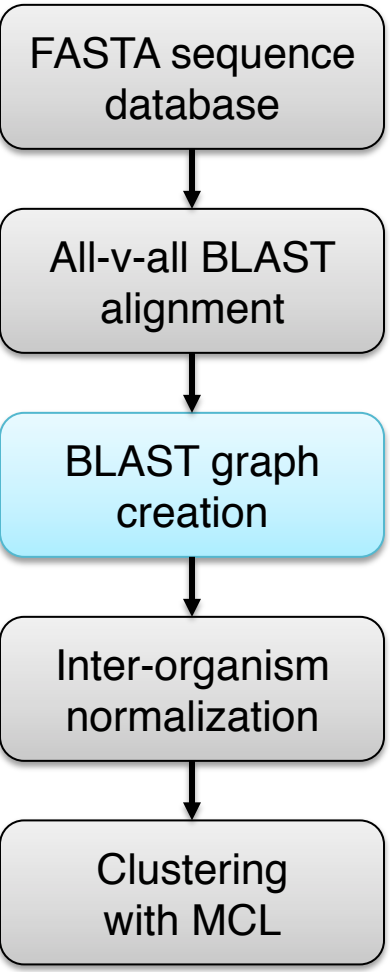
BLAST graph



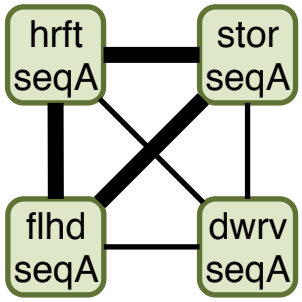
Orthologs



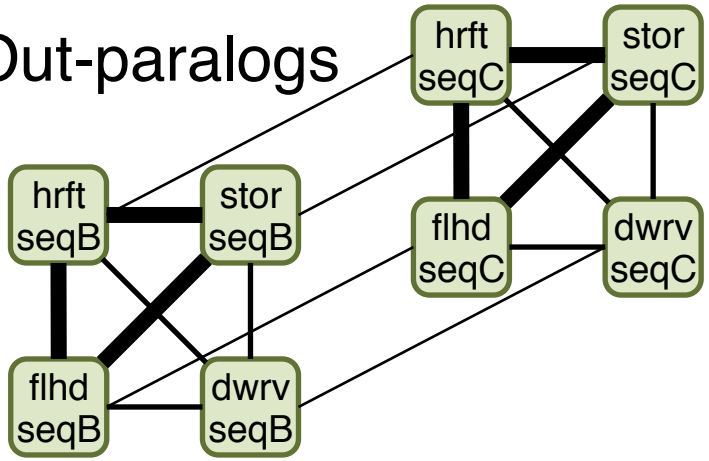
BLAST graph



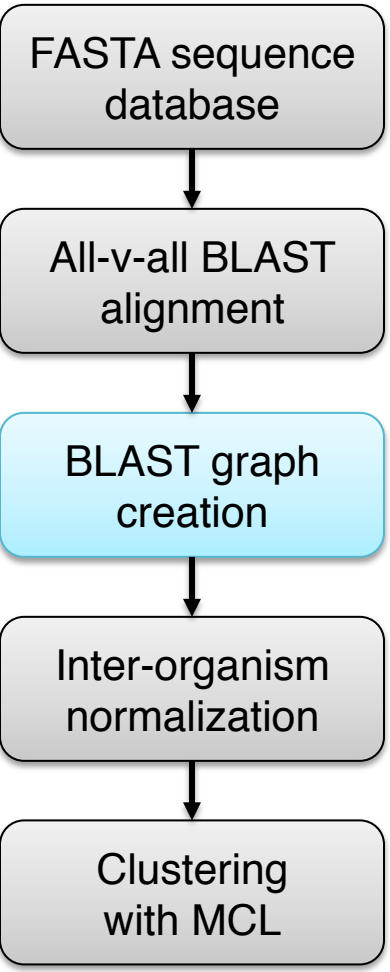
Orthologs

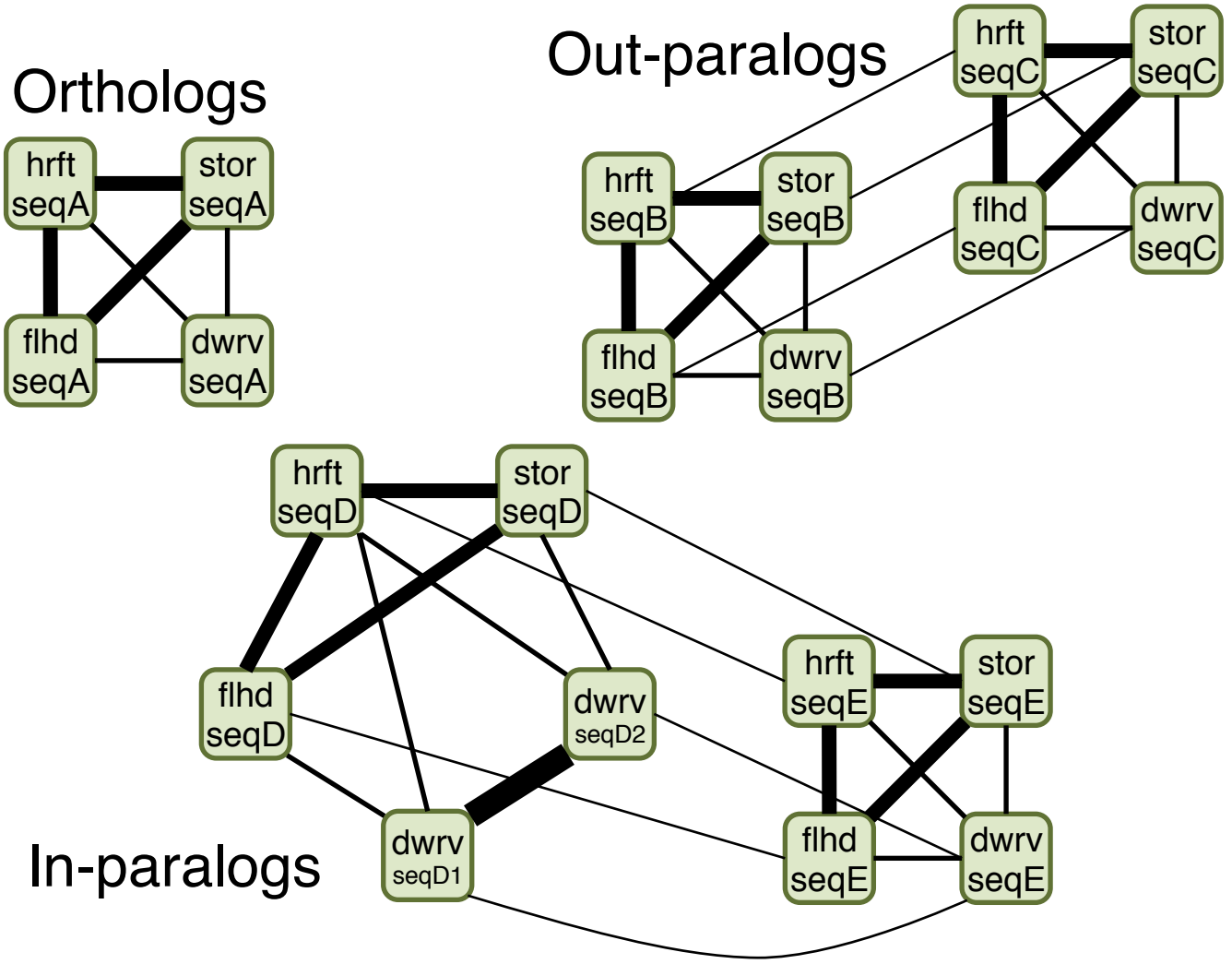


Out-paralogs

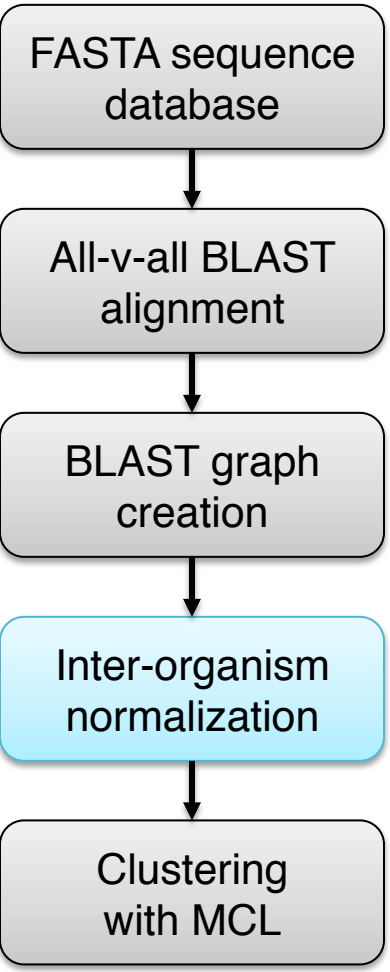


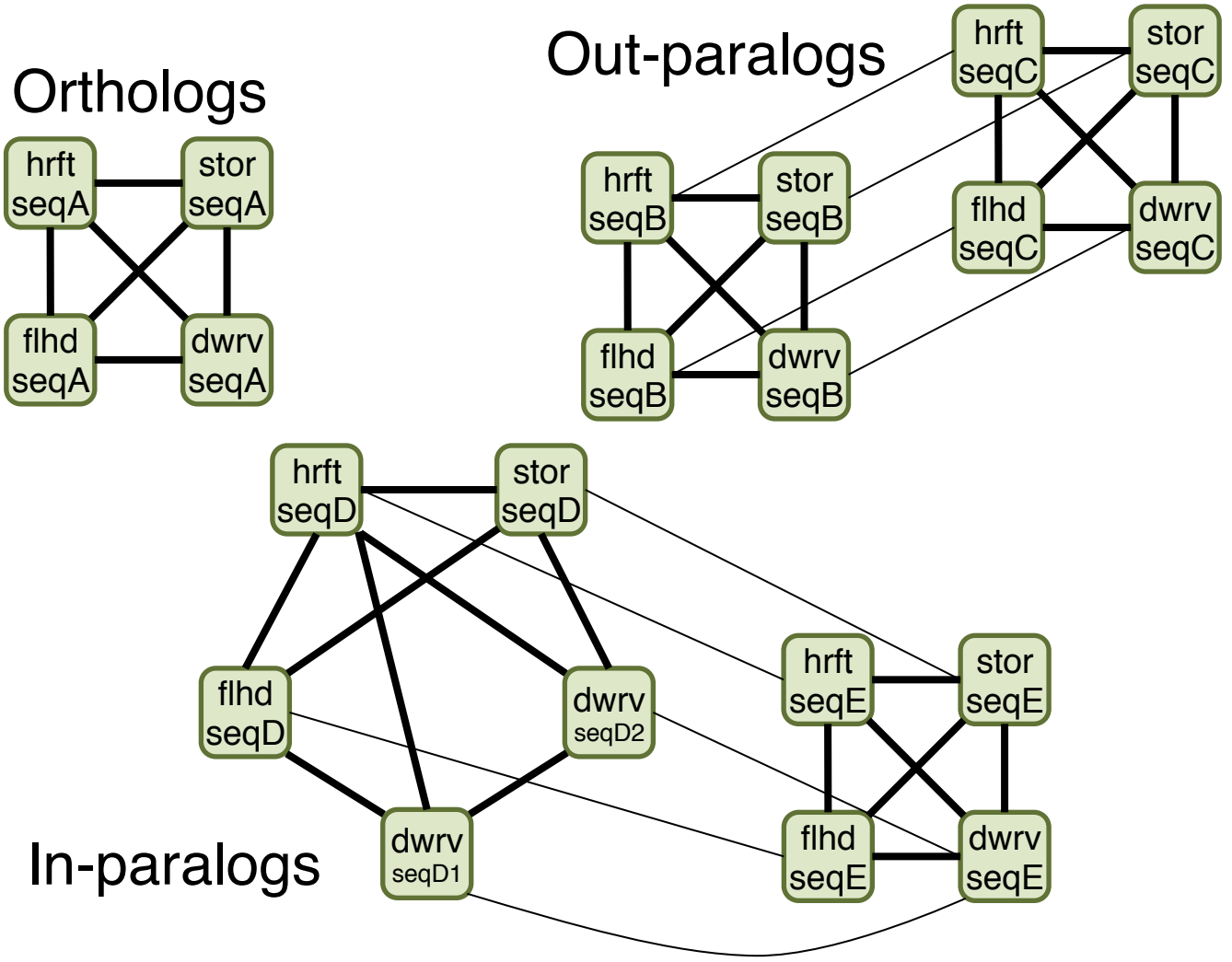
BLAST graph


FASTA sequence database
All-v-all BLAST alignment
BLAST graph creation
Inter-organism normalization
Clustering with MCL

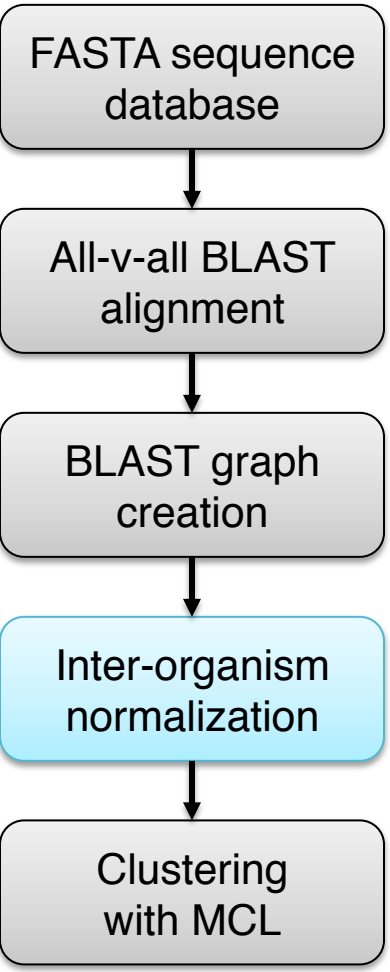

Orthologs
hrft seqA
stor seqA
flhd seqA
dwrv seqA
Out-paralogs
hrft seqB
stor seqB
flhd seqB
dwrv seqB
hrft seqC
stor seqC
flhd seqC
dwrv seqC
In-paralogs
hrft seqD
stor seqD
flhd seqD
dwrv seqD2
dwrv seqD1
hrft seqE
stor seqE
flhd seqE
dwrv seqE

Edge weight normalization





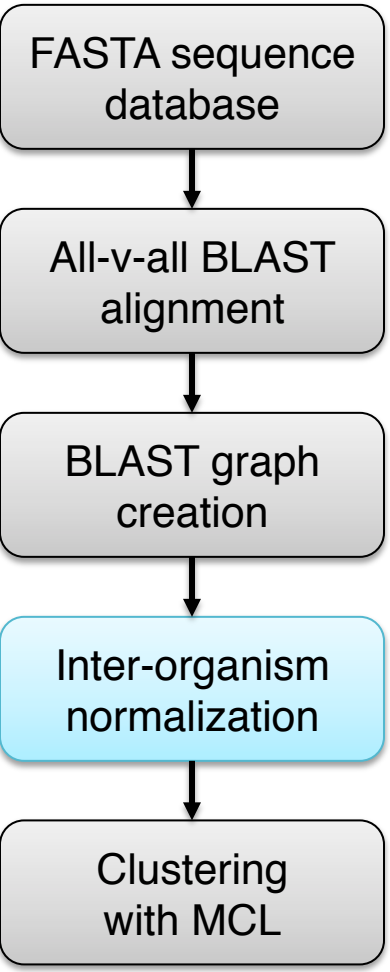
Edge weight normalization



$$w'_{A_i B_j} = w_{A_i B_j} \times \left(\frac{\bar{w}_G}{\bar{w}_{AB}} \right)$$

www.codecogs.com/latex/eqneditor.php

Edge weight normalization



$$w'_{A_iB_j} = w_{A_iB_j} \times \left(\frac{\bar{w}_G}{\bar{w}_{AB}} \right)$$

$w'_{A_iB_j}$ – normalized edge weight between sequence i in organism A and sequence j in organism B

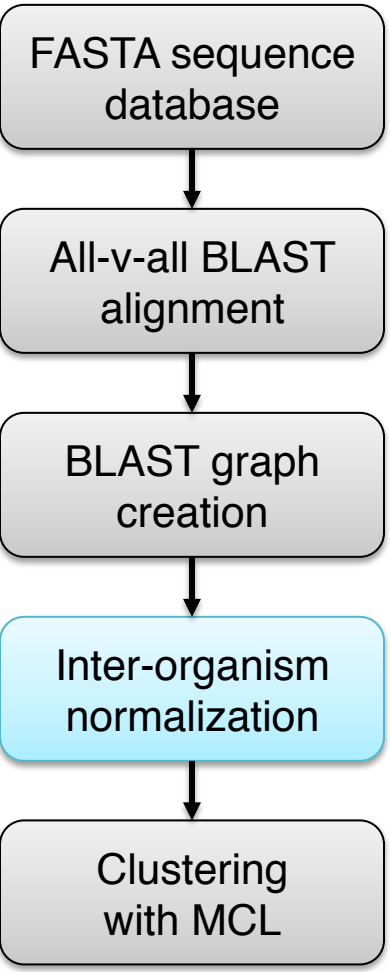
$w_{A_iB_j}$ – corresponding raw edge weight

$\bar{w}_G$ – average raw edge weight for the entire graph

$\bar{w}_{AB}$ – average raw edge weight between any pair of sequences in organisms A and B, respectively

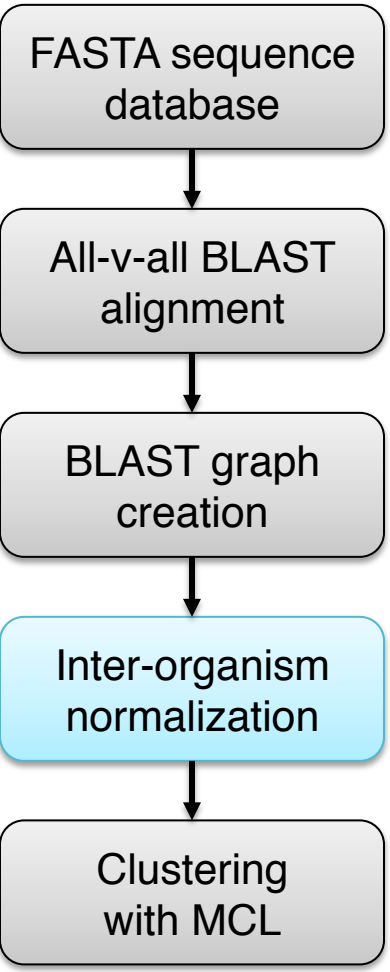
www.codecogs.com/latex/eqneditor.php

Normalized BLAST graph



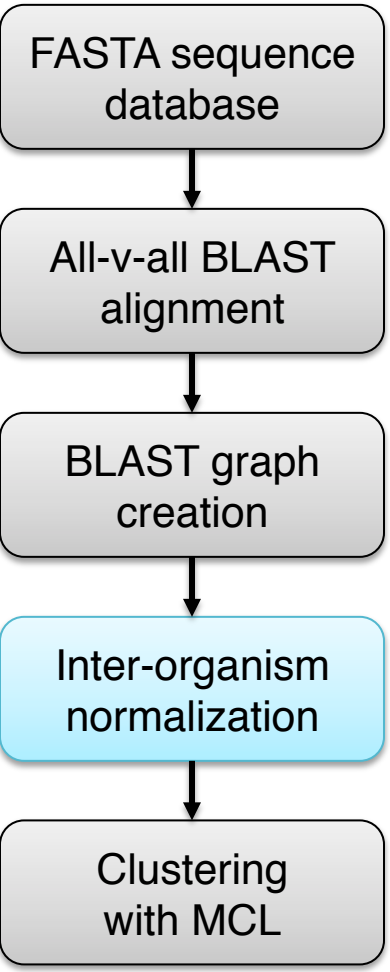
	hrft|seq0	stor|seq0	flhd|seq0	dwrv|seq0	hrft|seq1	stor|seq1	flhd|seq1
hrft|seq0	251	0	131	79	0	140	0
stor|seq0	0	165	0	0	0	96	137
flhd|seq0	131	0	282	0	119	95	147
dwrv|seq0	79	0	0	181	85	131	0
hrft|seq1	0	0	119	85	205	145	135
stor|seq1	140	96	95	131	145	253	0
flhd|seq1	0	137	147	0	135	0	238

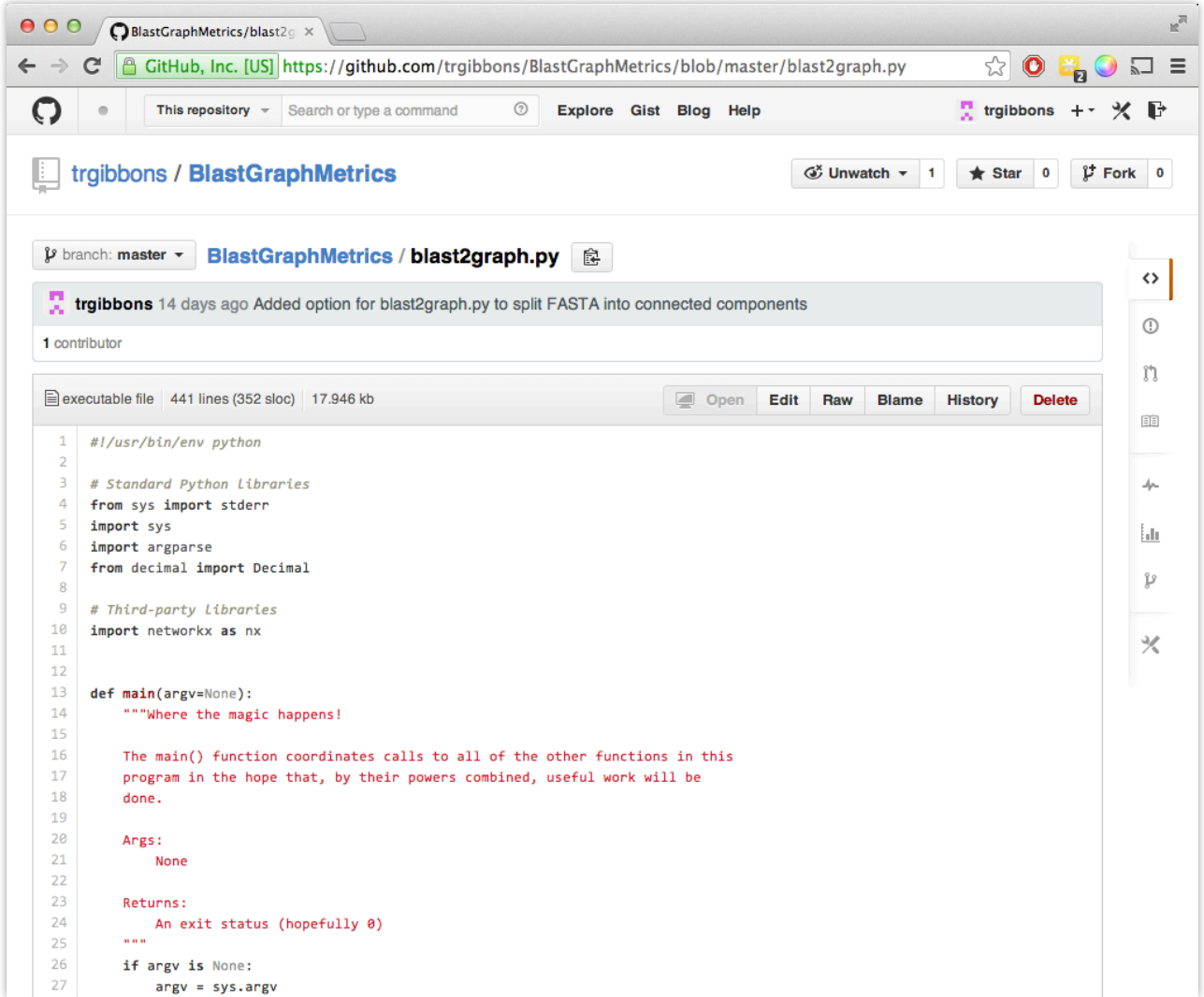
Normalized BLAST graph



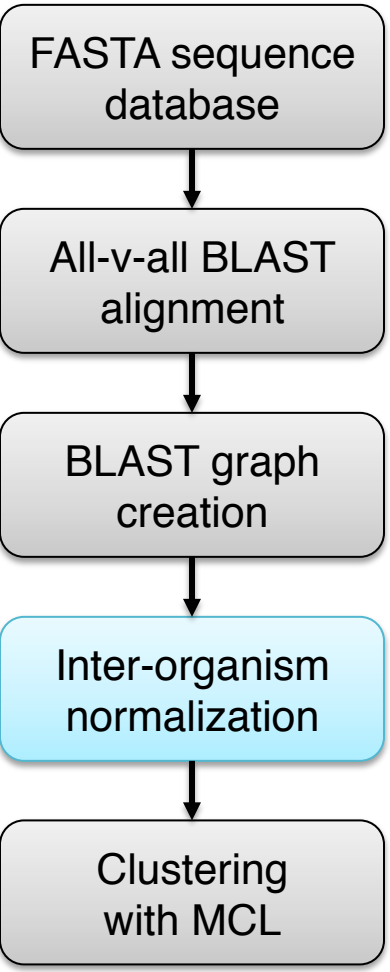
	hrft|seq0	stor|seq0	flhd|seq0	dwrv|seq0	hrft|seq1	stor|seq1	flhd|seq1
hrft|seq0	0	**0**	**109**	**99**	**0**	**127**	**0**
stor|seq0	0	0	**0**	**0**	**0**	**74**	**124**
flhd|seq0	131	0	0	**0**	**99**	**86**	**113**
dwrv|seq0	79	0	0	0	**106**	**164**	**0**
hrft|seq1	0	0	119	85	0	**132**	**112**
stor|seq1	140	96	95	131	145	0	**0**
flhd|seq1	0	137	147	0	135	0	0

Commands


FASTA sequence database
All-v-all BLAST alignment
BLAST graph creation
Inter-organism normalization
Clustering with MCL

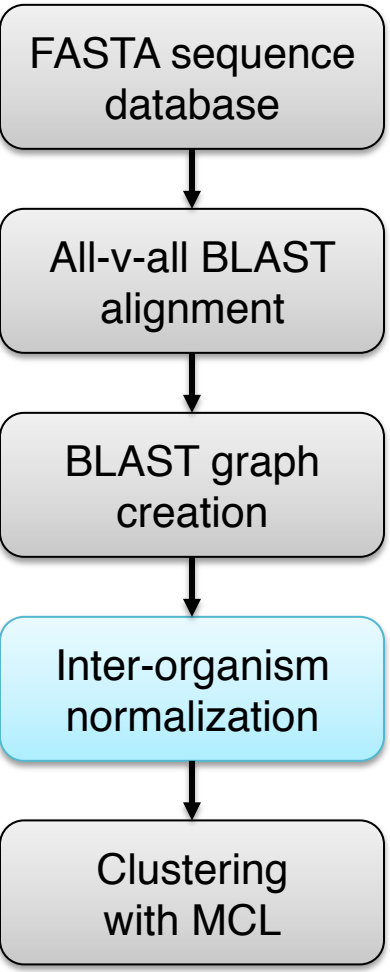

BlastGraphMetrics/blast2g
GitHub, Inc. [US] https://github.com/trgibbons/BlastGraphMetrics/blob/master/blast2graph.py
This repository
Search or type a command
Explore
Gist
Blog
Help
trgibbons
trgibbons / BlastGraphMetrics
Unwatch 1
Star 0
Fork 0
branch: master
BlastGraphMetrics / blast2graph.py
trgibbons 14 days ago Added option for blast2graph.py to split FASTA into connected components
1 contributor
executable file
441 lines (352 sloc)
17.946 kb
Open
Edit
Raw
Blame
History
Delete
#!/usr/bin/env python
# Standard Python libraries
from sys import stderr
import sys
import argparse
from decimal import Decimal
# Third-party libraries
import networkx as nx
def main(argv=None):
"""Where the magic happens!
The main() function coordinates calls to all of the other functions in this
program in the hope that, by their powers combined, useful work will be
done.
Args:
None
Returns:
An exit status (hopefully 0)
"""
if argv is None:
argv = sys.argv

Commands



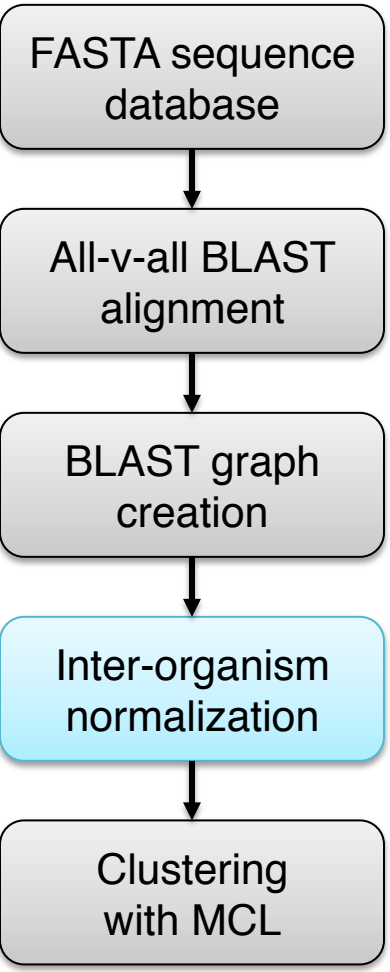

```
$ ./blast2graph.py -h
```

Commands



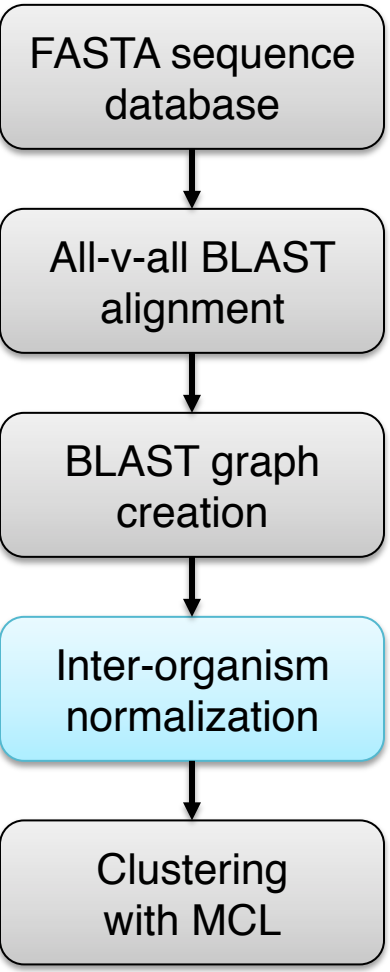

```
$ ./blast2graph.py -h
usage: blast2graph.py [-h] [--evcol EVCOL] [--bscol BSCOL] [--qlcol QLCOL]
                      [--slcol SLCOL] [--idchar IDCHAR] [--fasta FASTA] [-m]
                      blast out_pref

Generate a set of graphs from a tab-delimited BLASTP or BLASTN file such that
the first two columns contain the query and subject IDs, respectively, and the
last four columns contain, in order: E-value, bit score, query length, subject
length

positional arguments:
  blast             Tab-delimited BLAST file (comment lines are okay)
  out_pref          Prefix for the MCL-compatible "abc" graph files

optional arguments:
  -h, --help        show this help message and exit
  --evcol EVCOL     One-indexed column containing pairwise E-values (not
                    required if files include standard header lines) [def=11]
  --bscol BSCOL     One-indexed column containing pairwise bit scores (not
                    required if files include standard header lines) [def=12]
  --qlcol QLCOL     One-indexed column containing query lengths (not required
                    if files include standard header lines) [def=13]
  --slcol SLCOL     One-indexed column containing subject lengths (not required
                    if files include standard header lines) [def=14]
  --idchar IDCHAR   The character used to separate the organism ID from the
                    rest of the sequence header [def="|"]
  --fasta FASTA     FASTA file used to generate BLAST results, will be split
                    into connected components and reprinted, one file per
                    connected component
  -m, --merge       Merge sequences from a single organism when they have non-
                    overlapping alignments to the same target sequence
```

Commands



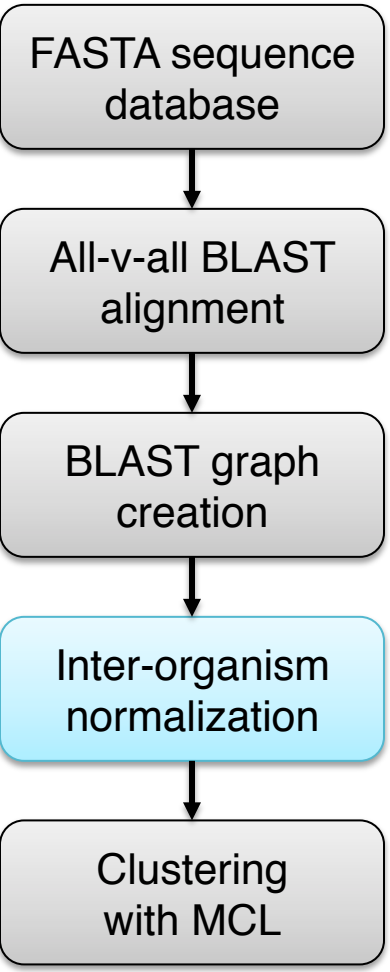

```
$ ./blast2graph.py -h
usage: blast2graph.py [-h] [--evcol EVCOL] [--bscol BSCOL] [--qlcol QLCOL]
                      [--slcol SLCOL] [--idchar IDCHAR] [--fasta FASTA] [-m]
                      blast out_pref

Generate a set of graphs from a tab-delimited BLASTP or BLASTN file such that
the first two columns contain the query and subject IDs, respectively, and the
last four columns contain, in order: E-value, bit score, query length, subject
length

positional arguments:
  blast             Tab-delimited BLAST file (comment lines are okay)
  out_pref          Prefix for the MCL-compatible "abc" graph files

optional arguments:
  -h, --help        show this help message and exit
  --evcol EVCOL     One-indexed column containing pairwise E-values (not
                    required if files include standard header lines) [def=11]
  --bscol BSCOL     One-indexed column containing pairwise bit scores (not
                    required if files include standard header lines) [def=12]
  --qlcol QLCOL     One-indexed column containing query lengths (not required
                    if files include standard header lines) [def=13]
  --slcol SLCOL     One-indexed column containing subject lengths (not required
                    if files include standard header lines) [def=14]
  --idchar IDCHAR   The character used to separate the organism ID from the
                    rest of the sequence header [def="|"]
  --fasta FASTA     FASTA file used to generate BLAST results, will be split
                    into connected components and reprinted, one file per
                    connected component
  -m, --merge       Merge sequences from a single organism when they have non-
                    overlapping alignments to the same target sequence
```

Commands




```
$ ./blast2graph.py db.blastp db
```

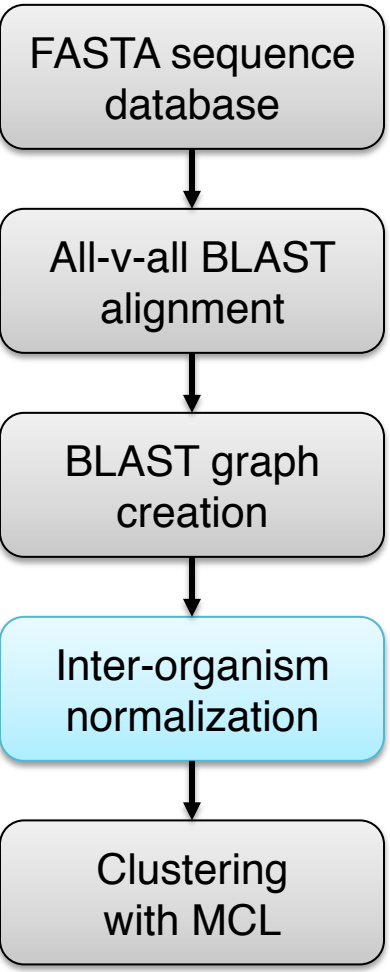
Commands




```
$ ./blast2graph.py db.blastp db
```

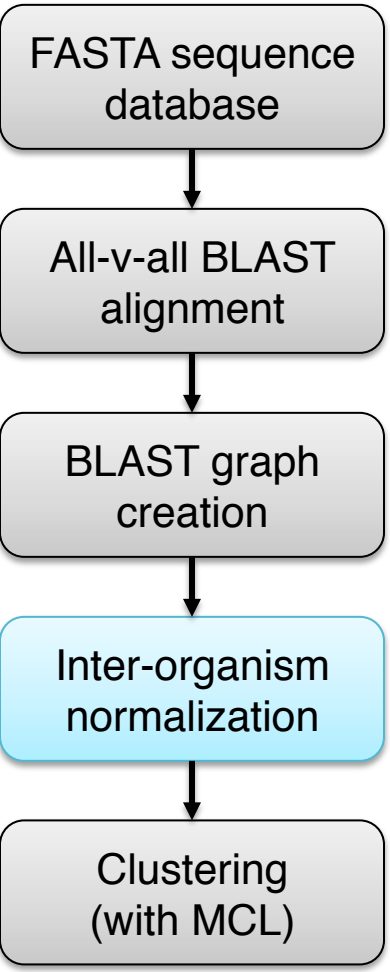
```
$ ls -1 db_*.abc
  db_raw_bit.abc
  db_raw_bpr.abc
  db_raw_bsr.abc
  db_raw_pev.abc
  db_nrm_bit.abc
  db_nrm_bpr.abc
  db_nrm_bsr.abc
  db_nrm_pev.abc
```

Commands



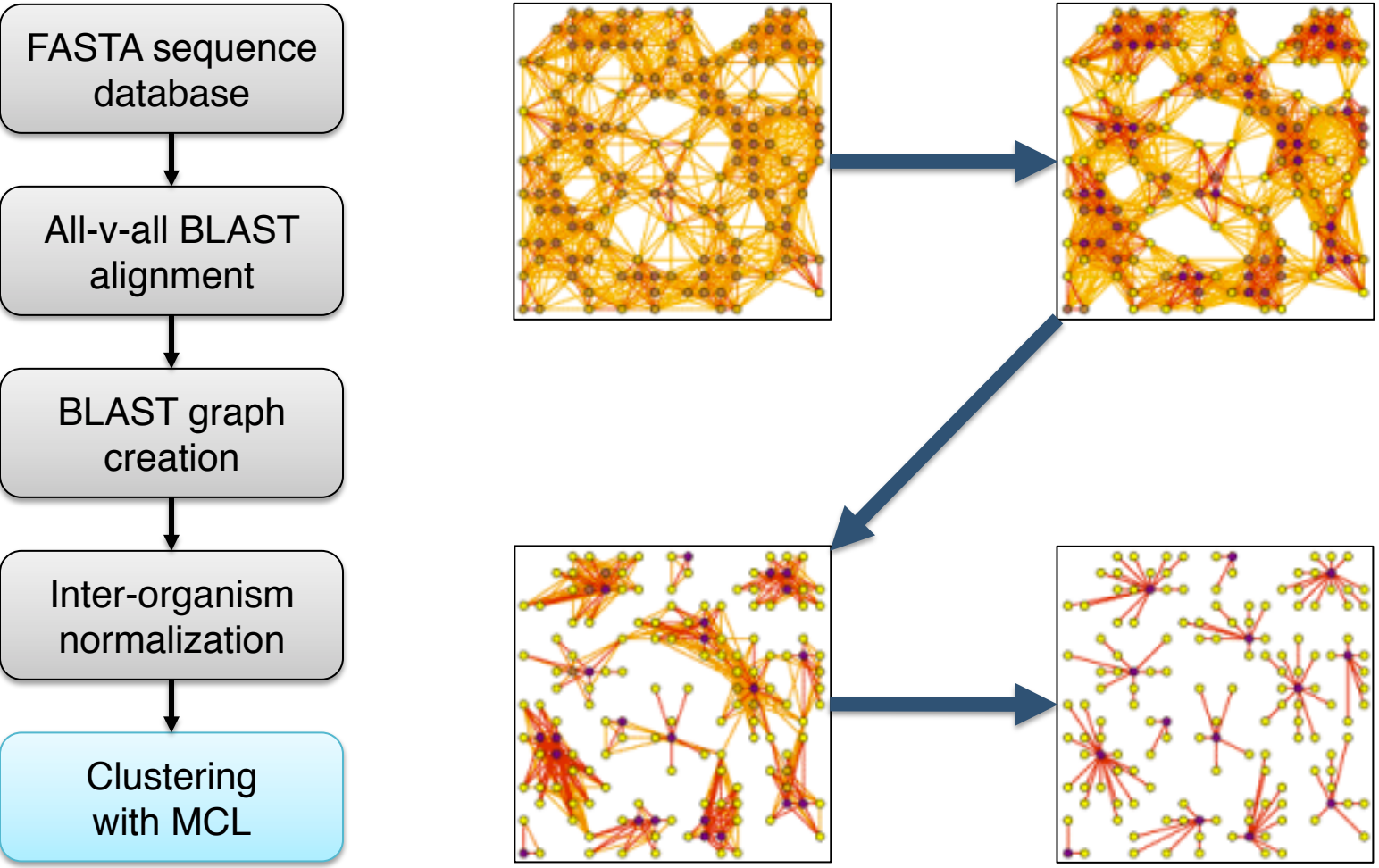

```
$ cat db_raw_bit.abc
  hrft|seq0    stor|seq89    87.8
  hrft|seq0    hrft|seq3     84.3
  hrft|seq0    flhd|seq29    82.0
  hrft|seq0    hrft|seq2     80.1
  hrft|seq0    stor|seq15    78.2
  hrft|seq0    flhd|seq397   76.3
  hrft|seq0    hrft|seq12    72.4
  hrft|seq0    flhd|seq19    72.0
  hrft|seq0    stor|seq6     69.7
  hrft|seq0    hrft|seq59    67.8
  hrft|seq0    flhd|seq76    65.5
  hrft|seq0    stor|seq33    59.3
  hrft|seq0    stor|seq0     57.4
  hrft|seq0    hrft|seq456   57.0
  hrft|seq0    hrft|seq937   54.7
  ...
```

Normalized BLAST graph

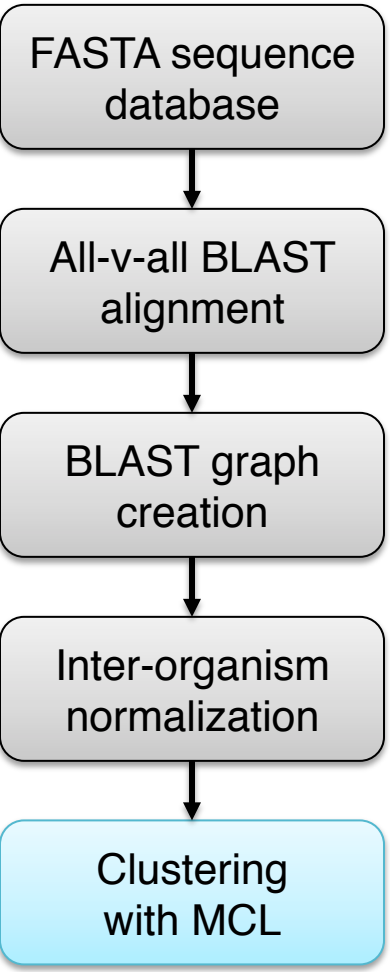


	hrft|seq0	stor|seq0	flhd|seq0	dwrv|seq0	hrft|seq1	stor|seq1	flhd|seq1
hrft|seq0	0	**0**	**109**	**99**	**0**	**127**	**0**
stor|seq0	0	0	**0**	**0**	**0**	**74**	**124**
flhd|seq0	131	0	0	**0**	**99**	**86**	**113**
dwrv|seq0	79	0	0	0	**106**	**164**	**0**
hrft|seq1	0	0	119	85	0	**132**	**112**
stor|seq1	140	96	95	131	145	0	**0**
flhd|seq1	0	137	147	0	135	0	0

Markov **CL**ustering (MCL)

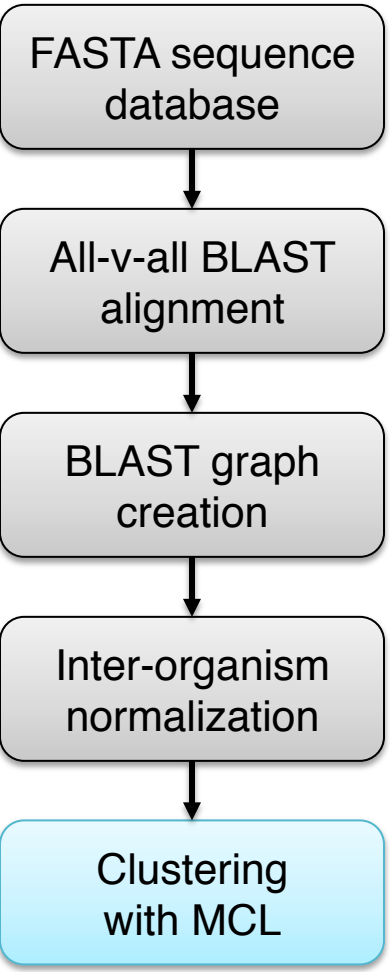


Commands




```
$ mcl db_nrm_bit.abc \
      --abc \
      -I 1.2 \
      -o db_nrm_bit_I12.mcl
```

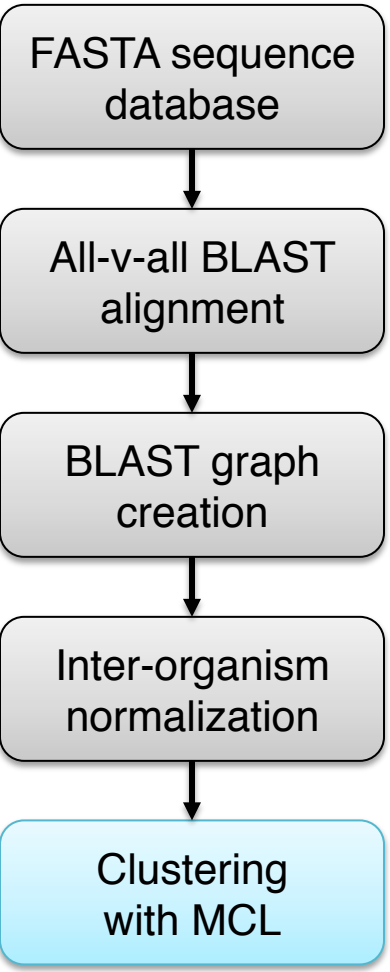
Commands




```
$ mcl db_nrm_bit.abc \
      --abc \
      -I 1.2 \
      -o db_nrm_bit_I12.mcl
```

```
$ for I in $(seq -w 11 25)
> do
>     i1=${I:0:1}
>     i2=${I:1}
>     mcl db_nrm_bit.abc \
>         --abc \
>         -I ${i1}.${i2} \
>         -o db_nrm_bit_I${I}.mcl
> done
```

Commands




```
$ cat db_nrm_bit_I12.mcl
  dwrv|seqD1  dwrv|seqD2  hrft|seqD  flhd|seqD  stor|seqD
  hrft|seqA   stor|seqA   flhd|seqA  dwrv|seqA
  flhd|seqB   stor|seqB   hrft|seqB  dwrv|seqA
  stor|seqC   hrft|seqC   flhd|seqC  dwrv|seqA
  hrft|seqE   stor|seqE   dwrv|seqE  flhd|seqA
```

Teaser for part II

FASTA sequence database

All-v-all BLAST alignment

BLAST graph creation

Inter-organism normalization

Clustering with MCL

Bonus!

Bit Score

$$S' = \frac{\lambda S - \ln(K)}{\ln(2)}$$

Bit Per Residue

$$BPR = \frac{S'}{\min(l_i, l_j)}$$

Bit Score Ratio

$$BSR = \frac{S'}{\min(SBS_i, SBS_j)}$$

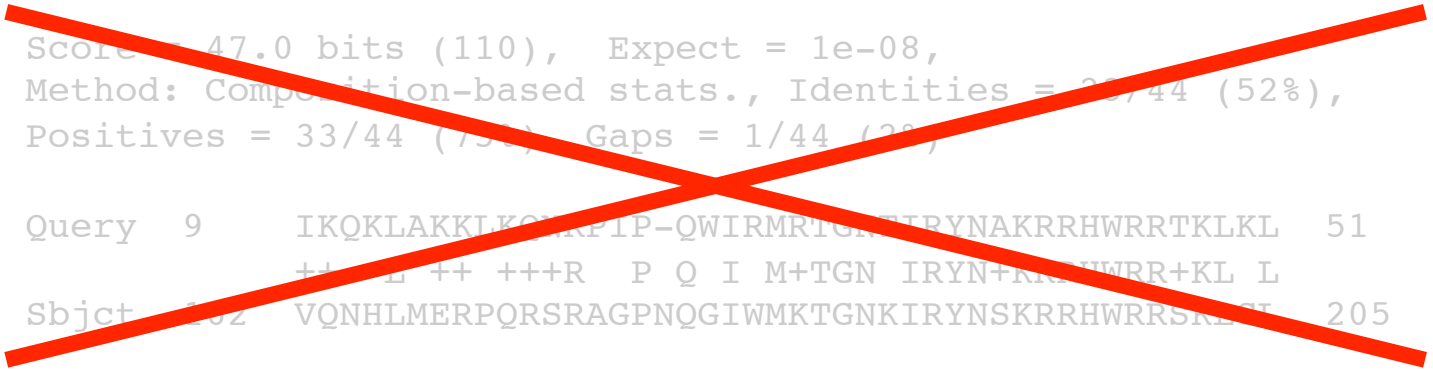
"E"(xpect) value

$$E = \frac{l_i L}{2^{S'}}$$

Teaser for part II

- FASTA sequence database
- All-v-all BLAST alignment
- BLAST graph creation
- Inter-organism normalization
- Clustering with MCL

Bonus!



```
# BLASTP 2.2.28+
# Query: hrft|seq0
# Database: one_db_to_rule_them_all.fasta
# Fields: query id, subject id, % identity, alignment length, mismatches, gap opens, q. start, q. end, s. start, ...
# 21 hits found
```

hrft|seq0	hrft|seq0	100.00	52	0	0	1	52	1	52	6e-31	102	52	52
hrft|seq0	stor|seq89	84.31	51	8	0	1	51	1	51	5e-25	87.8	52	51
hrft|seq0	hrft|seq3	76.47	51	12	0	1	51	1	51	1e-23	84.3	52	51
hrft|seq0	flhd|seq29	78.85	52	11	0	1	52	1	52	7e-23	82.0	52	52
hrft|seq0	hrft|seq2	76.47	51	12	0	1	51	1	51	4e-22	80.1	52	51
hrft|seq0	stor|seq15	74.51	51	13	0	1	51	1	51	2e-21	78.2	52	51
hrft|seq0	flhd|seq397	88.24	51	6	0	1	51	1	51	1e-20	76.3	52	51
hrft|seq0	hrft|seq12	69.23	52	16	0	1	52	1	52	4e-19	72.4	52	52
hrft|seq0	flhd|seq19	79.07	43	9	0	9	51	2	44	5e-19	72.0	52	44
hrft|seq0	stor|seq6	80.49	41	8	0	9	49	2	42	5e-18	69.7	52	44
hrft|seq0	hrft|seq59	74.51	51	13	0	1	51	1	51	2e-17	67.8	52	51
hrft|seq0	flhd|seq76	82.69	52	9	0	1	52	1	52	2e-16	65.5	52	52
hrft|seq0	stor|seq33	65.96	47	13	1	5	51	48	91	8e-14	59.3	52	91
hrft|seq0	stor|seq0	77.78	36	8	0	6	41	153	188	1e-12	57.4	52	202
hrft|seq0	hrft|seq456	76.32	38	7	1	6	43	380	415	8e-12	57.0	52	415
hrft|seq0	hrft|seq937	76.67	30	7	0	22	51	543	572	8e-11	54.7	52	572
hrft|seq0	stor|seq52	75.00	28	7	0	24	51	248	275	2e-09	49.7	52	275
hrft|seq0	dwrv|seq16	49.02	51	26	0	1	51	1	51	4e-09	46.2	52	52
hrft|seq0	dwrv|seq150	52.27	44	20	1	9	51	162	205	1e-08	47.0	52	205
hrft|seq0	flhd|seq47	76.92	26	6	0	26	51	94	119	3e-08	44.3	52	119
hrft|seq0	hrft|seq0	73.08	26	7	0	26	51	281	306	7e-08	45.8	52	379

Teaser for part II

FASTA sequence database

All-v-all BLAST alignment

BLAST graph creation

Inter-organism normalization

Clustering with MCL

Bonus!

Bit Score

$$S' = \frac{\lambda S - \ln(K)}{\ln(2)}$$

Bit Per Residue

$$BPR = \frac{S'}{\min(l_i, l_j)}$$

Bit Score Ratio

$$BSR = \frac{S'}{\min(SBS_i, SBS_j)}$$

"E"(xpect) value

$$E = \frac{l_i L}{2^{S'}}$$

Teaser for part II

FASTA sequence database

All-v-all BLAST alignment

BLAST graph creation

Inter-organism normalization

Clustering with MCL

Bonus!

Bit Score

$$S' = \frac{\lambda S - \ln(K)}{\ln(2)}$$

Bit Per Residue

$$BPR = \frac{S'}{\min(l_i, l_j)}$$

Bit Score Ratio

$$BSR = \frac{S'}{\min(SBS_i, SBS_j)}$$

p(E-value)

$$p(E) = -\log_{10}\left(\frac{l_i L}{2^{S'}}\right)$$

Teaser for part II

FASTA sequence database

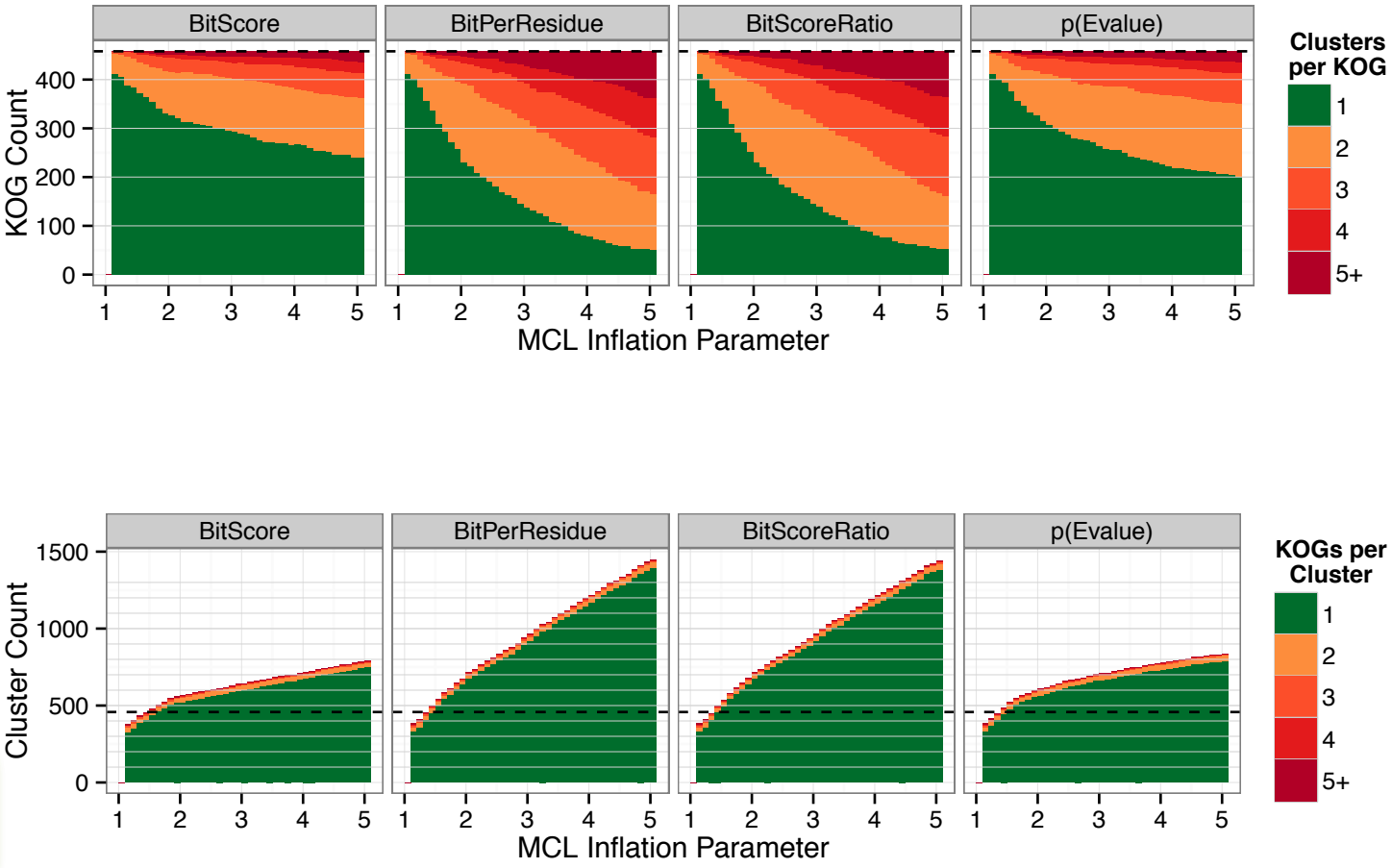
All-v-all BLAST alignment

BLAST graph creation

Inter-organism normalization

Clustering with MCL

Bonus!



Tree views

Part II:
Effects of graph weighting metrics and the inflation parameter

Ted Gibbons

Graph Weighting Metrics

Bit Score

$$S' = \frac{\lambda S - \ln(K)}{\ln(2)}$$

Bit Per Residue

$$BPR = \frac{S'}{\min(l_i, l_j)}$$

Bit Score Ratio

$$BSR = \frac{S'}{\min(SBS_i, SBS_j)}$$

"E"(xpect) value

$$E = \frac{l_i L}{2^{S'}}$$

Graph Weighting Metrics

~~Score = 47.0 bits (110), Expect = 1e-08,~~
~~Method: Compositional-based stats., Identities = 23/44 (52%),~~
~~Positives = 33/44 (75%), Gaps = 1/44 (2%)~~

~~Query 9 IKQKLAKKLKQNKPIP-QWIRMRTGNIRYNAKRRHWRRTKLKL 51~~
~~+ L ++ +++R P Q I M+TGN IRYN+KRRHWRR+KL L~~
~~Sbjct 162 VQNHLMERPQRSRAGPNQGIWMKTGNKIRYNSKRRHWRRSKLSI 205~~

```
# BLASTP 2.2.28+
# Query: hrft|seq0
# Database: one_db_to_rule_them_all.fasta
# Fields: query id, subject id, % identity, alignment length, mismatches, gap opens, q. start, q. end, s. start, ...
# 21 hits found
```

hrft|seq0	hrft|seq0	100.00	52	0	0	1	52	1	52	6e-31	102	52	52
hrft|seq0	stor|seq89	84.31	51	8	0	1	51	1	51	5e-25	87.8	52	51
hrft|seq0	hrft|seq3	76.47	51	12	0	1	51	1	51	1e-23	84.3	52	51
hrft|seq0	flhd|seq29	78.85	52	11	0	1	52	1	52	7e-23	82.0	52	52
hrft|seq0	hrft|seq2	76.47	51	12	0	1	51	1	51	4e-22	80.1	52	51
hrft|seq0	stor|seq15	74.51	51	13	0	1	51	1	51	2e-21	78.2	52	51
hrft|seq0	flhd|seq397	88.24	51	6	0	1	51	1	51	1e-20	76.3	52	51
hrft|seq0	hrft|seq12	69.23	52	16	0	1	52	1	52	4e-19	72.4	52	52
hrft|seq0	flhd|seq19	79.07	43	9	0	9	51	2	44	5e-19	72.0	52	44
hrft|seq0	stor|seq6	80.49	41	8	0	9	49	2	42	5e-18	69.7	52	44
hrft|seq0	hrft|seq59	74.51	51	13	0	1	51	1	51	2e-17	67.8	52	51
hrft|seq0	flhd|seq76	82.69	52	9	0	1	52	1	52	2e-16	65.5	52	52
hrft|seq0	stor|seq33	65.96	47	13	1	5	51	48	91	8e-14	59.3	52	91
hrft|seq0	stor|seq0	77.78	36	8	0	6	41	153	188	1e-12	57.4	52	202
hrft|seq0	hrft|seq456	76.32	38	7	1	6	43	380	415	8e-12	57.0	52	415
hrft|seq0	hrft|seq937	76.67	30	7	0	22	51	543	572	8e-11	54.7	52	572
hrft|seq0	stor|seq52	75.00	28	7	0	24	51	248	275	2e-09	49.7	52	275
hrft|seq0	dwrv|seq16	49.02	51	26	0	1	51	1	51	4e-09	46.2	52	52
hrft|seq0	dwrv|seq150	52.27	44	20	1	9	51	162	205	1e-08	47.0	52	205
hrft|seq0	flhd|seq47	76.92	26	6	0	26	51	94	119	3e-08	44.3	52	119
hrft|seq0	hrft|seq0	73.08	26	7	0	26	51	281	306	7e-08	45.8	52	379

Graph Weighting Metrics

Bit Score

$$S' = \frac{\lambda S - \ln(K)}{\ln(2)}$$

Bit Per Residue

$$BPR = \frac{S'}{\min(l_i, l_j)}$$

Bit Score Ratio

$$BSR = \frac{S'}{\min(SBS_i, SBS_j)}$$

"E"(xpect) value

$$E = \frac{l_i L}{2^{S'}}$$

Graph Weighting Metrics

Bit Score

$$S' = \frac{\lambda S - \ln(K)}{\ln(2)}$$

Bit Per Residue

$$BPR = \frac{S'}{\min(l_i, l_j)}$$

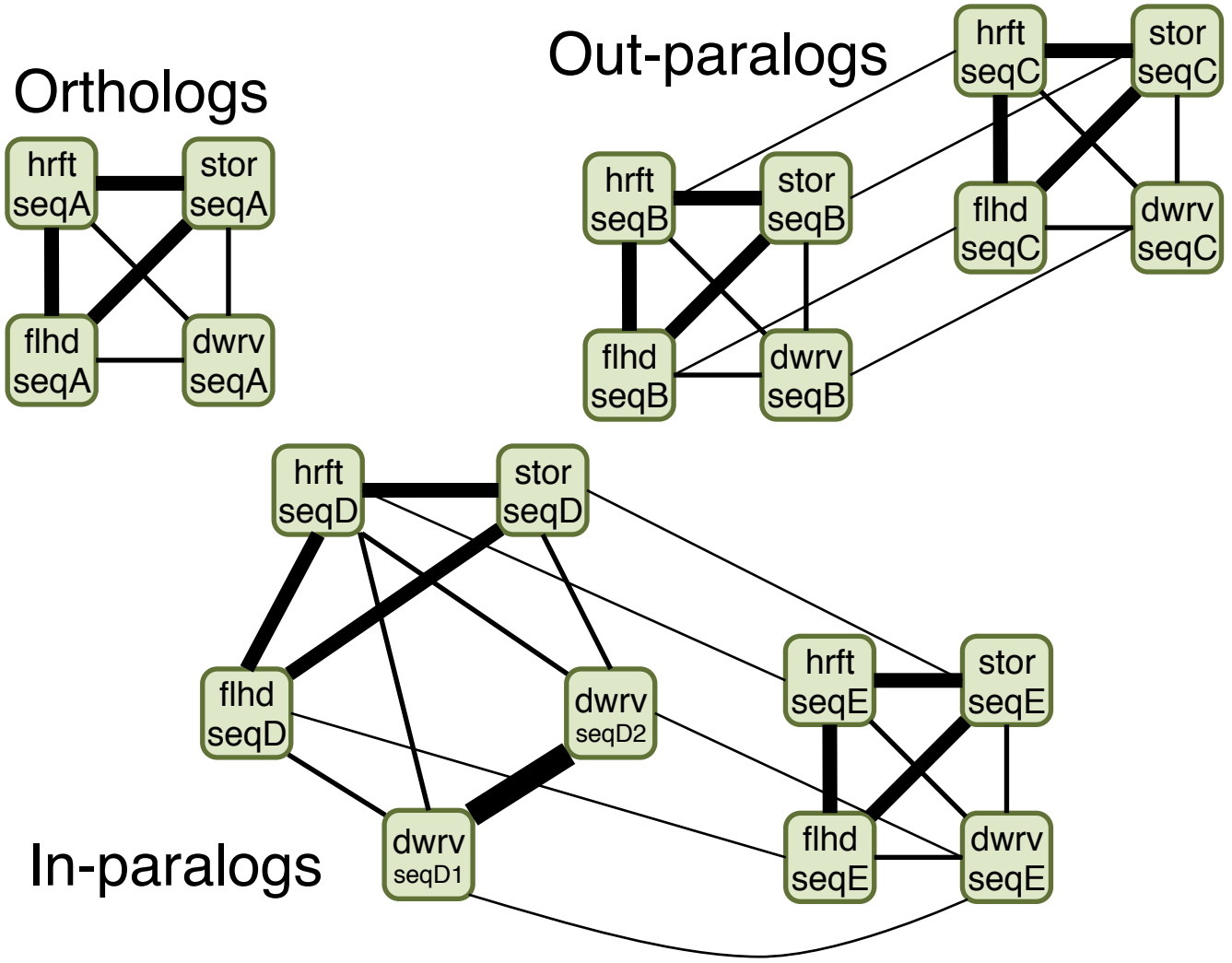
Bit Score Ratio

$$BSR = \frac{S'}{\min(SBS_i, SBS_j)}$$

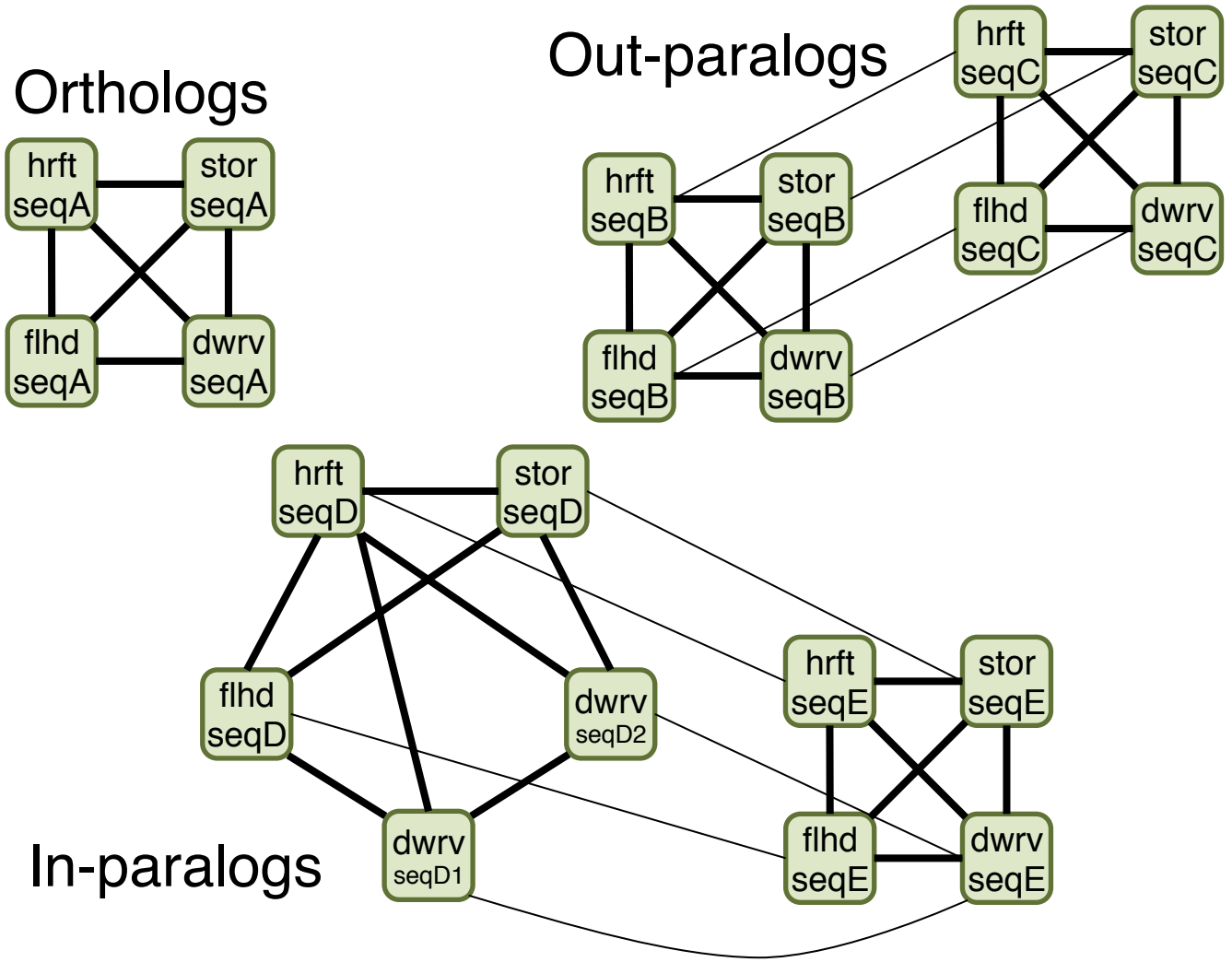
p(E-value)

$$p(E) = -\log_{10}\left(\frac{l_i L}{2^{S'}}\right)$$

Edge Weight Normalization



Edge Weight Normalization



Edge weight normalization

$$w'_{A_i B_j} = w_{A_i B_j} \times \left(\frac{\bar{w}_G}{\bar{w}_{AB}} \right)$$

$w'_{A_i B_j}$ – normalized edge weight between sequence i in organism A and sequence j in organism B

$w_{A_i B_j}$ – corresponding raw edge weight

$\bar{w}_G$ – average raw edge weight for the entire graph

$\bar{w}_{AB}$ – average raw edge weight between any pair of sequences in organisms A and B, respectively

www.codecogs.com/latex/eqneditor.php

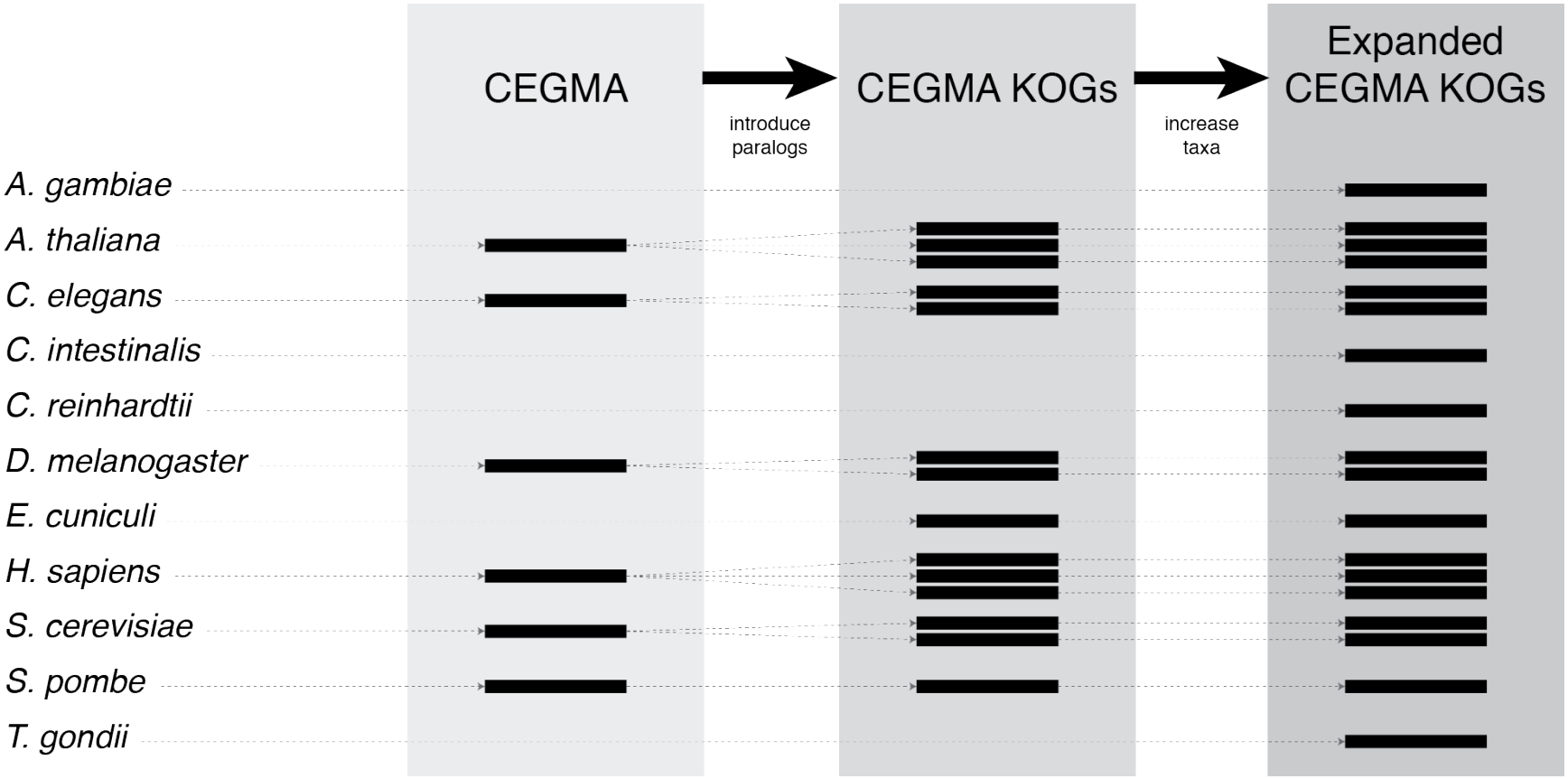
MCL Inflation Parameter

7.2 – **The effect of inflation on cluster granularity**

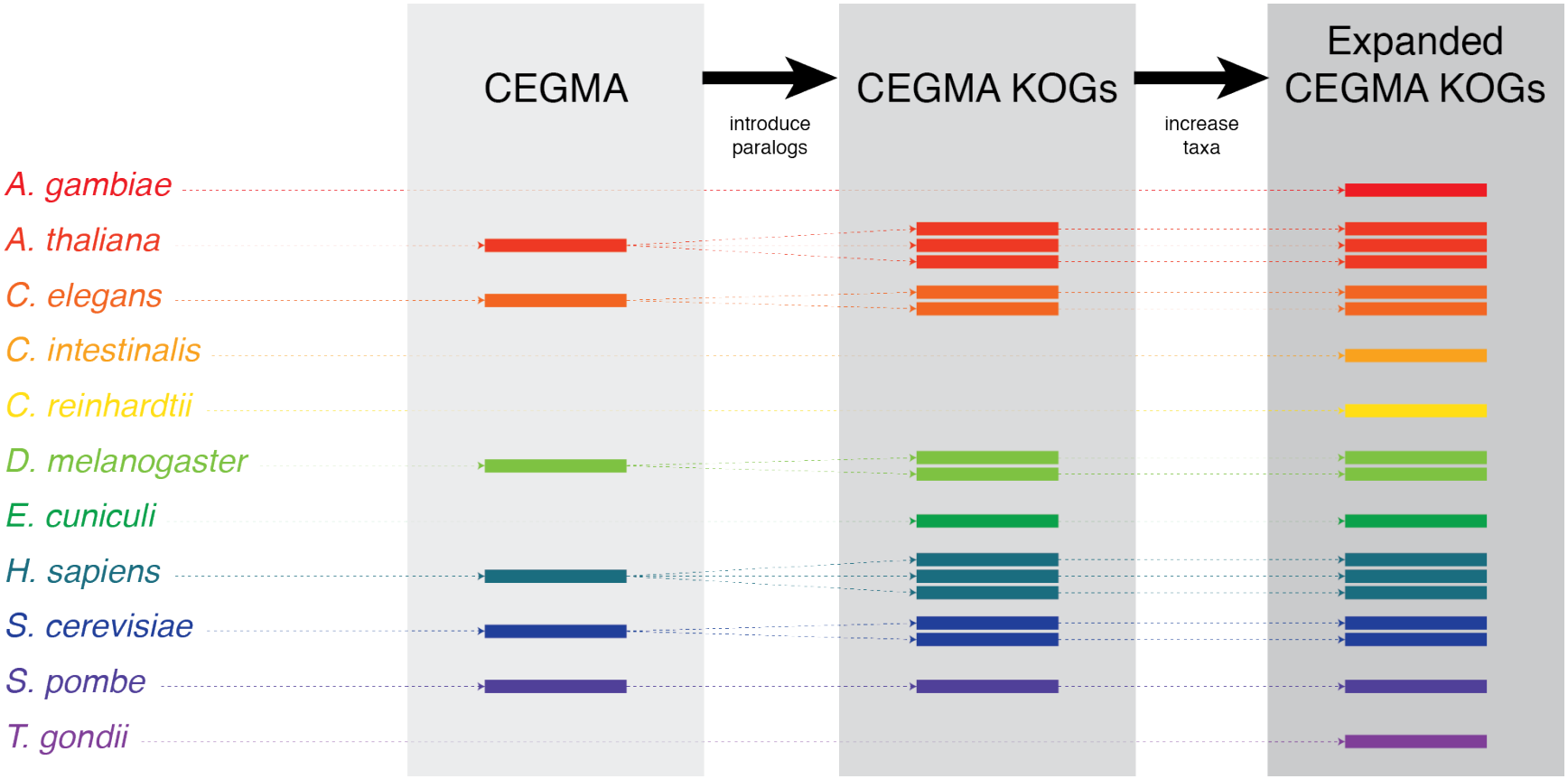
The main handle for changing inflation is the **-I** option. This is also *the* principal handle for regulating cluster granularity. Unless you are mangling huge graphs it could be the only **mcl** option you ever need besides the output redirection option **-o**.

Increasing the value of **-I** will increase cluster granularity. Conceivable values are from 1.1 to 10.0 or so, but the range of suitable values will certainly depend on your input graph. For many graphs, 1.1 will be far too low, and for many other graphs, 8.0 will be far too high. You will have to find the right value or range of values by experimenting, using your judgment, and using measurement tools such as clm dist and clm info. A good set of values to start with is 1.4, 2 and 6.

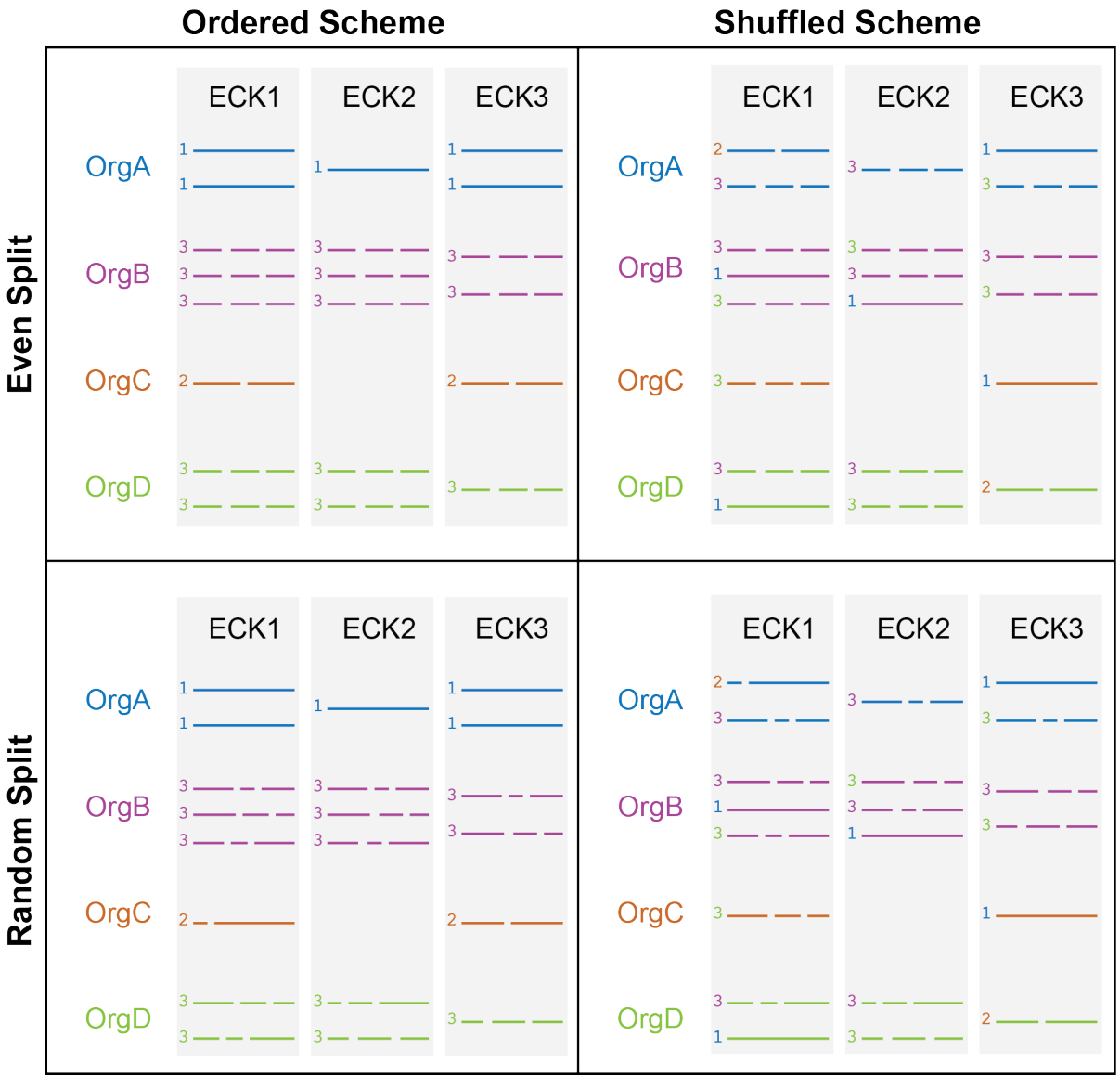
Test Database



Test Database

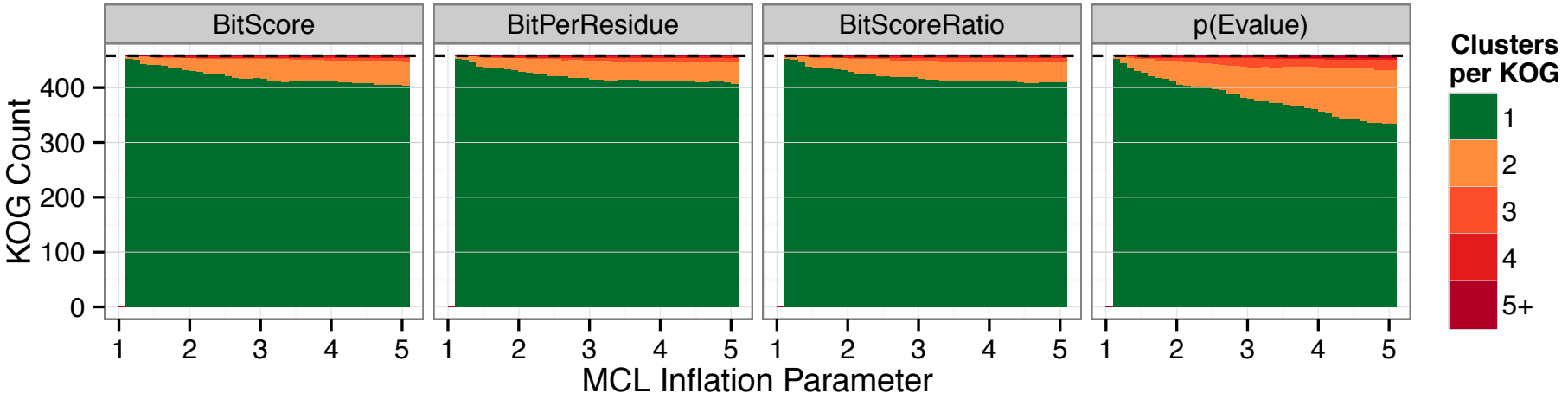


Fragmentation Scheme: **1323**



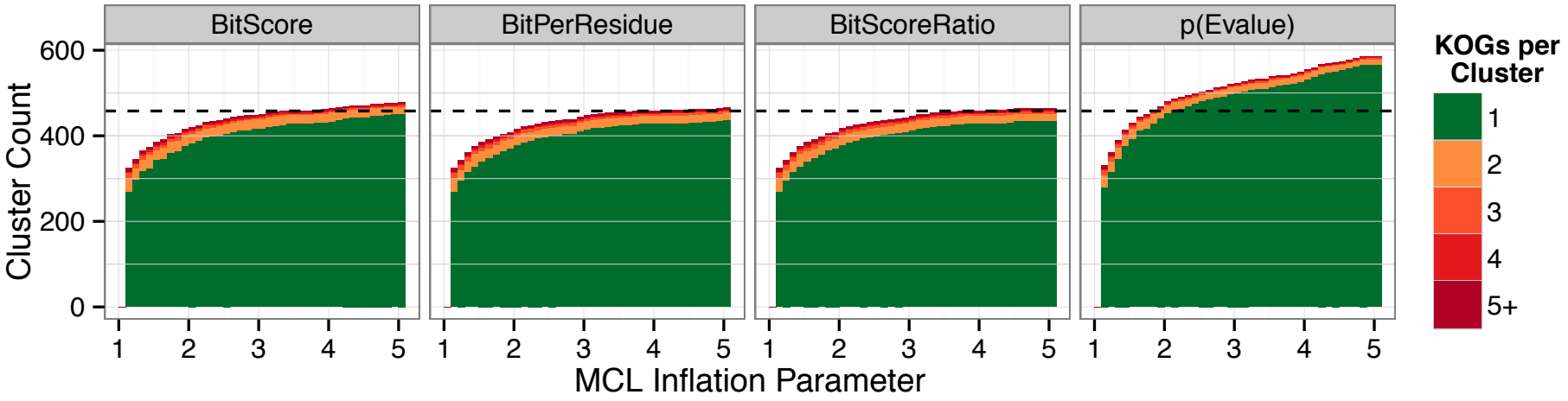
Sensitivity

eck_1/ord_evn/1e−5/nrm/eck_11111111111_ord_evn_1e−5_nrm_clusters_per_kog_summary.Rtab



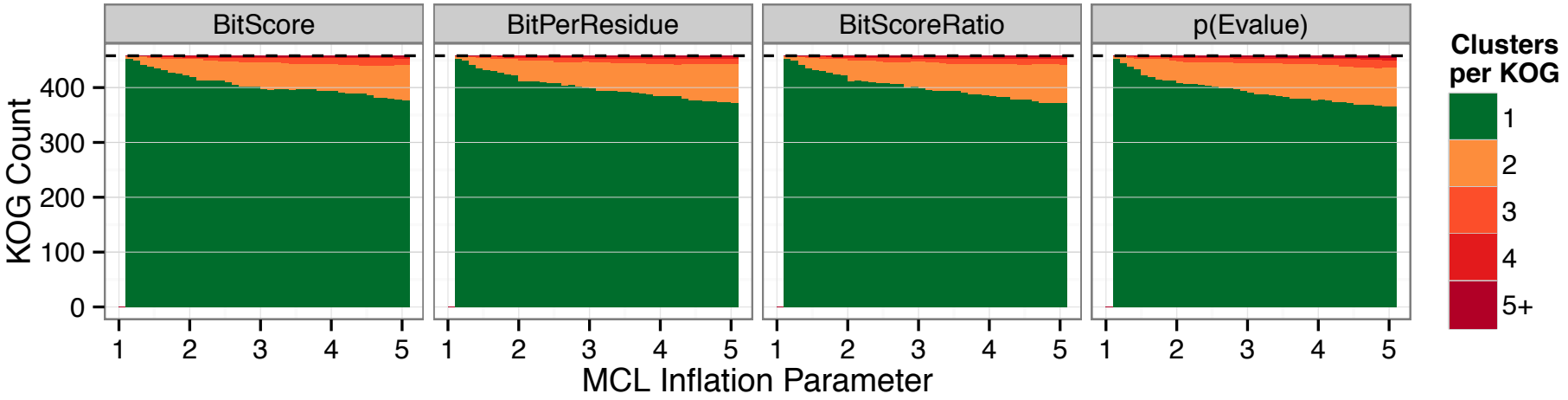
Specificity

eck_1/ord_evn/1e−5/nrm/eck_11111111111_ord_evn_1e−5_nrm_kogs_per_cluster_summary.Rtab



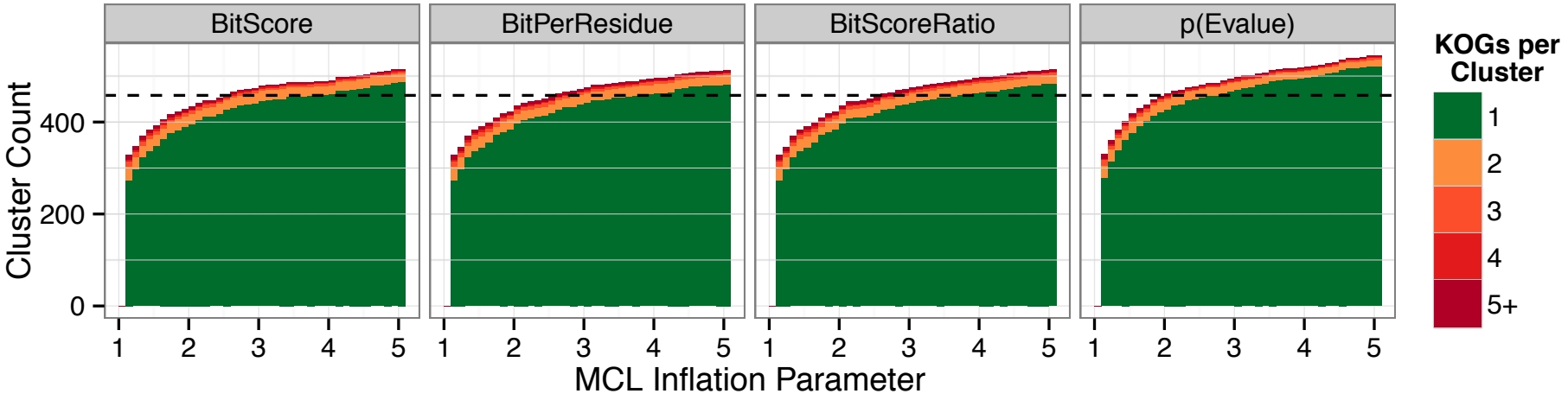
Sensitivity

eck_1/ord_evn/1e−5/raw/eck_11111111111_ord_evn_1e−5_raw_clusters_per_kog_summary.Rtab



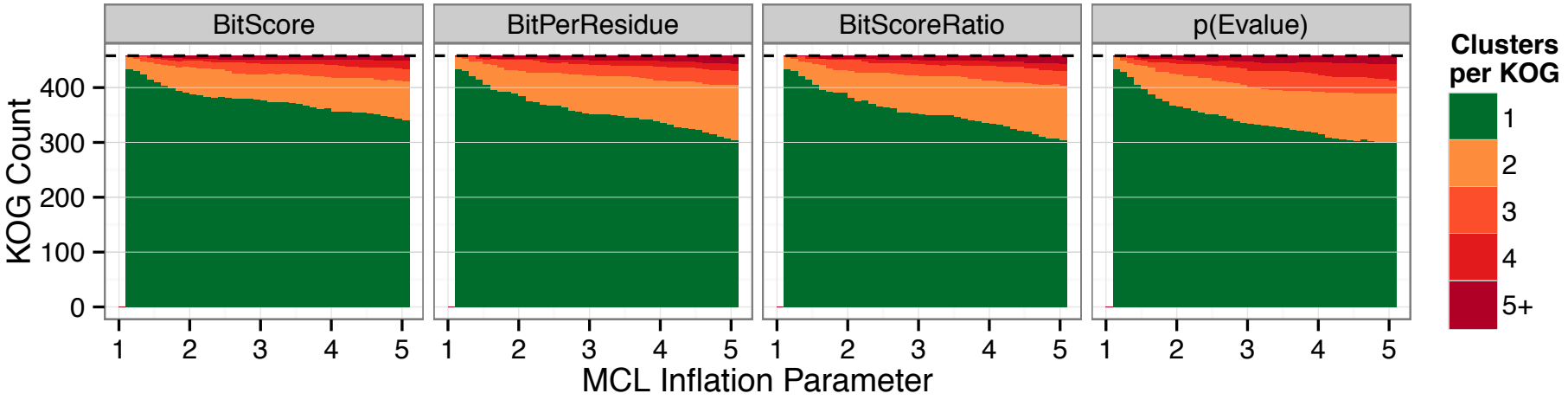
Specificity

eck_1/ord_evn/1e−5/raw/eck_11111111111_ord_evn_1e−5_raw_kogs_per_cluster_summary.Rtab



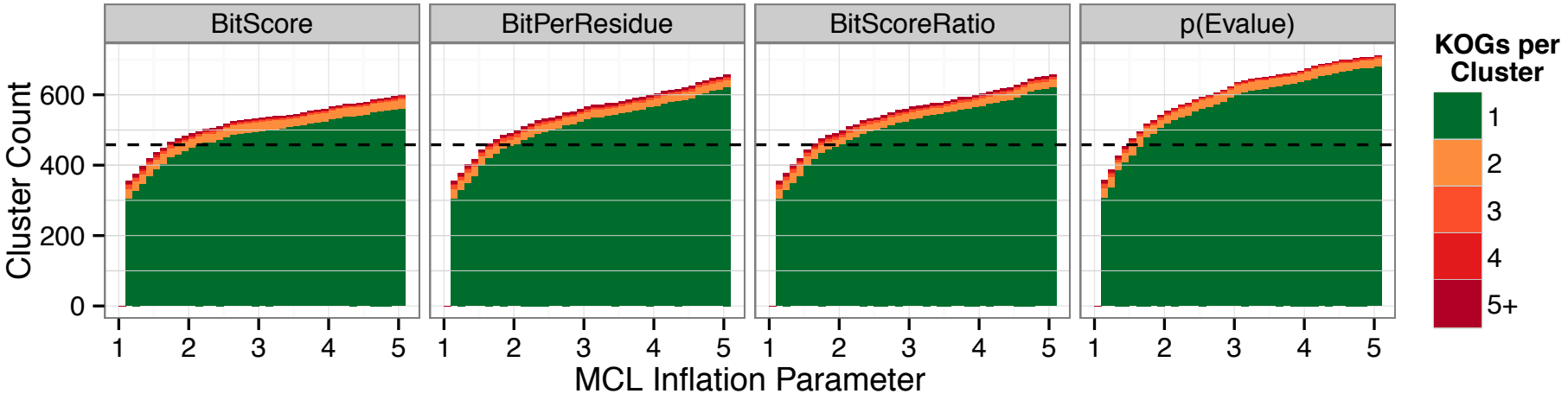
Sensitivity

eck_12/shf_evn/1e−5/nrm/eck_12121212121_shf_evn_1e−5_nrm_clusters_per_kog_summary.Rtab



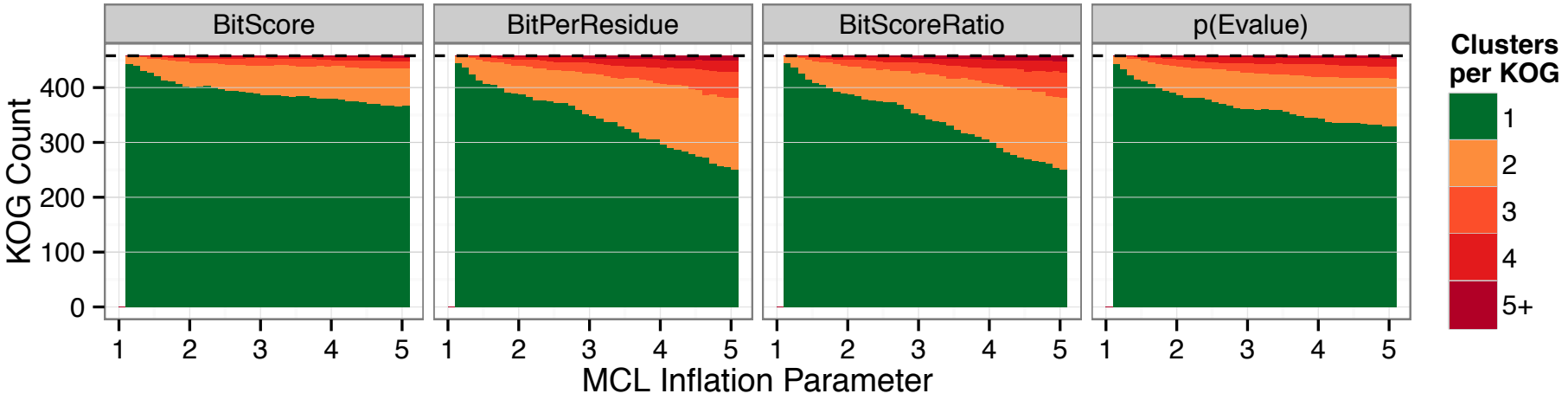
Specificity

eck_12/shf_evn/1e−5/nrm/eck_12121212121_shf_evn_1e−5_nrm_kogs_per_cluster_summary.Rtab



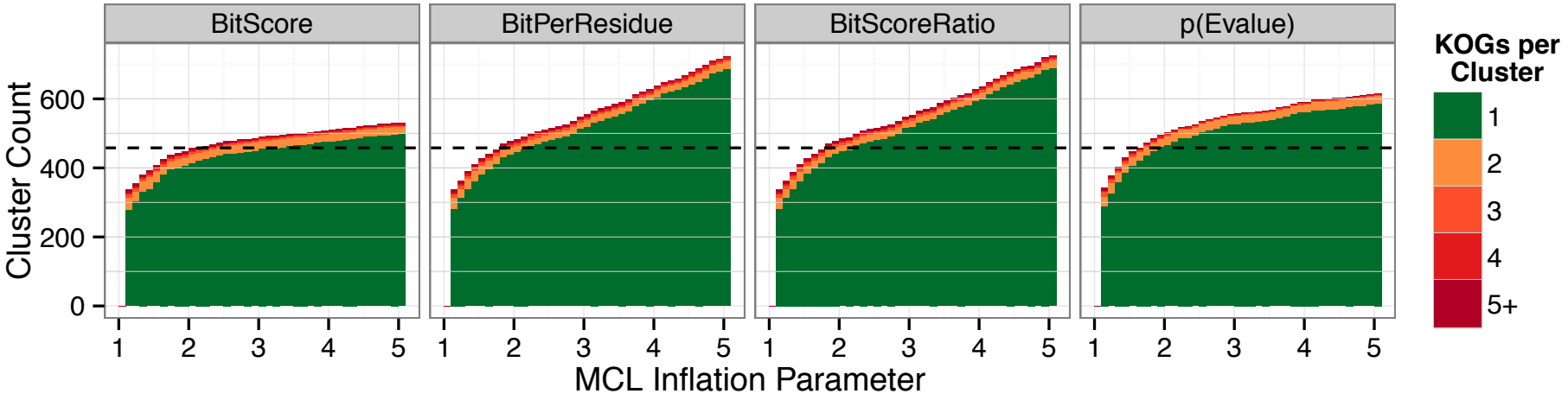
Sensitivity

eck_12/shf_rnd/1e−5/nrm/eck_12121212121_shf_rnd_1e−5_nrm_clusters_per_kog_summary.Rtab



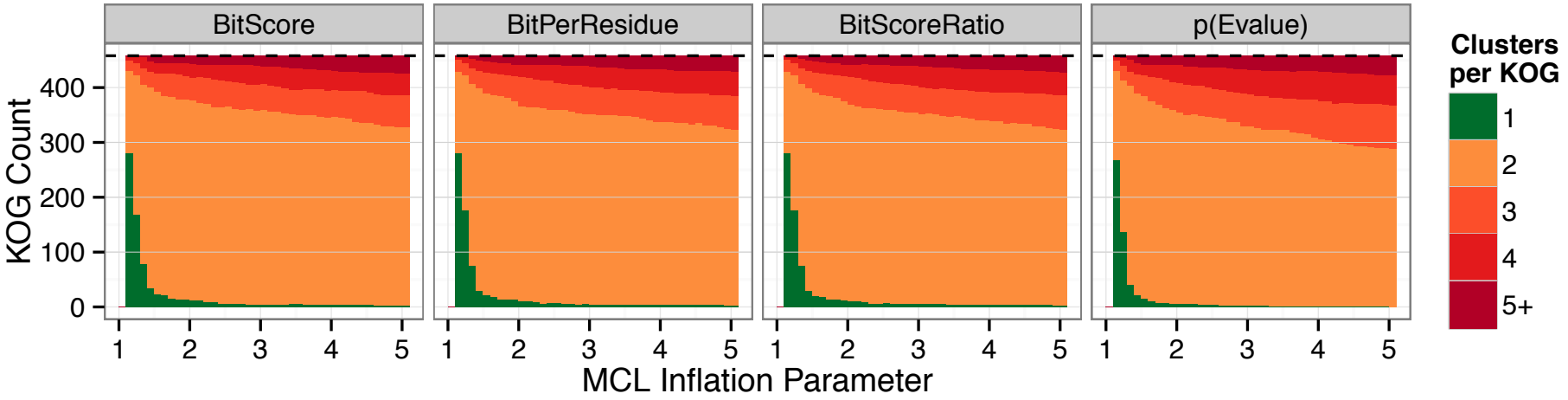
Specificity

eck_12/shf_rnd/1e−5/nrm/eck_12121212121_shf_rnd_1e−5_nrm_kogs_per_cluster_summary.Rtab



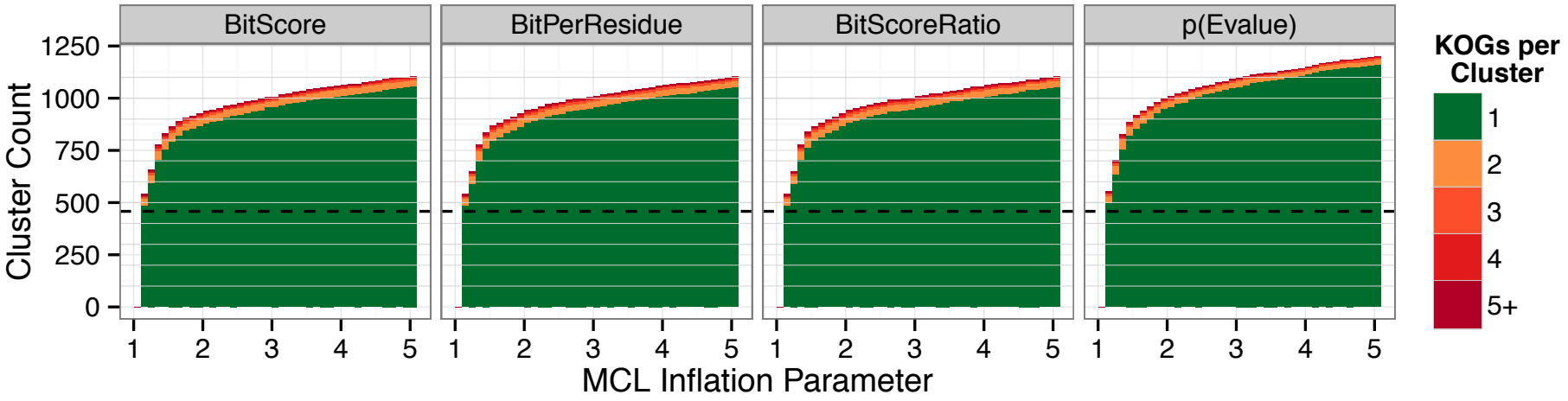
Sensitivity

eck_2/shf_evn/1e−5/nrm/eck_22222222222_shf_evn_1e−5_nrm_clusters_per_kog_summary.Rtab



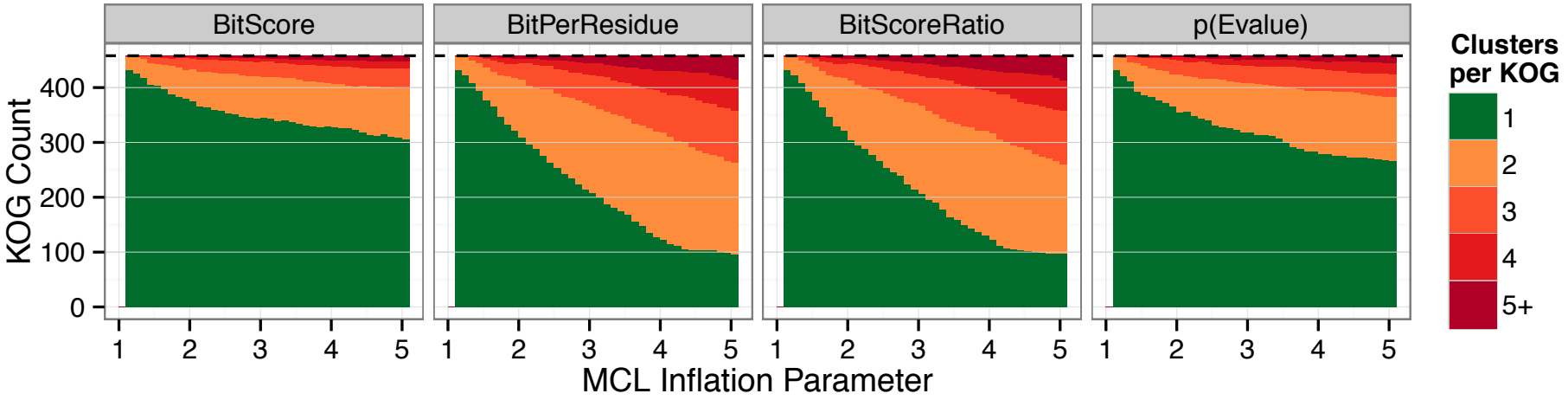
Specificity

eck_2/shf_evn/1e−5/nrm/eck_22222222222_shf_evn_1e−5_nrm_kogs_per_cluster_summary.Rtab



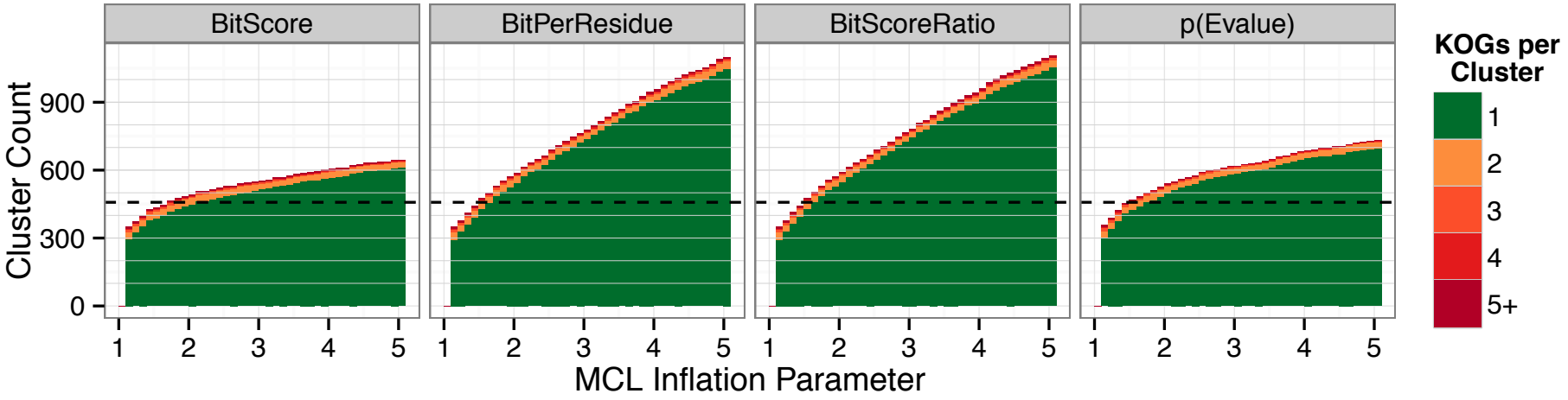
Sensitivity

eck_2/shf_rnd/1e−5/nrm/eck_22222222222_shf_rnd_1e−5_nrm_clusters_per_kog_summary.Rtab



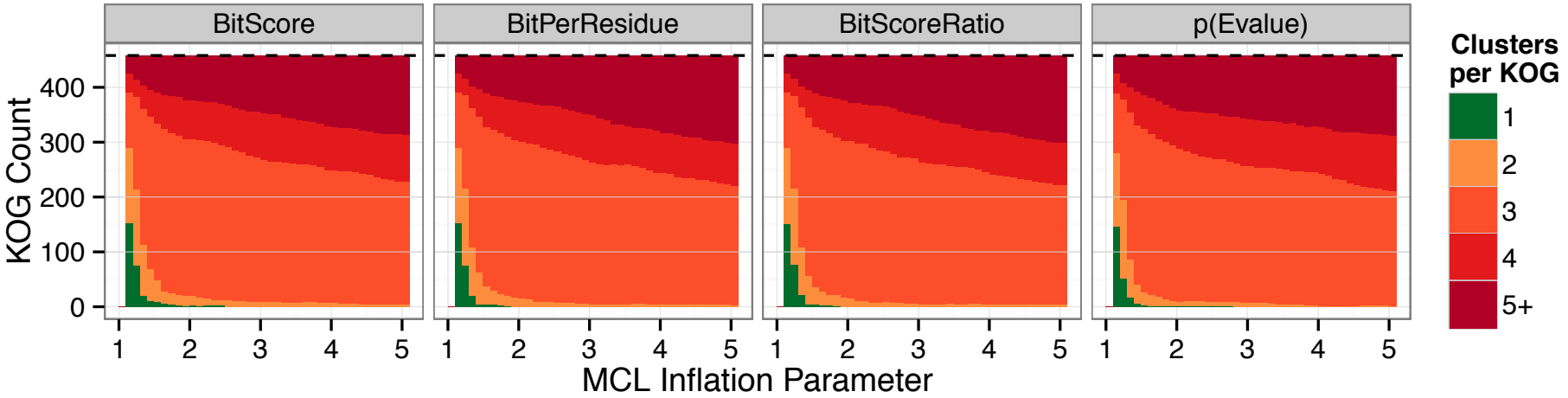
Specificity

eck_2/shf_rnd/1e−5/nrm/eck_22222222222_shf_rnd_1e−5_nrm_kogs_per_cluster_summary.Rtab



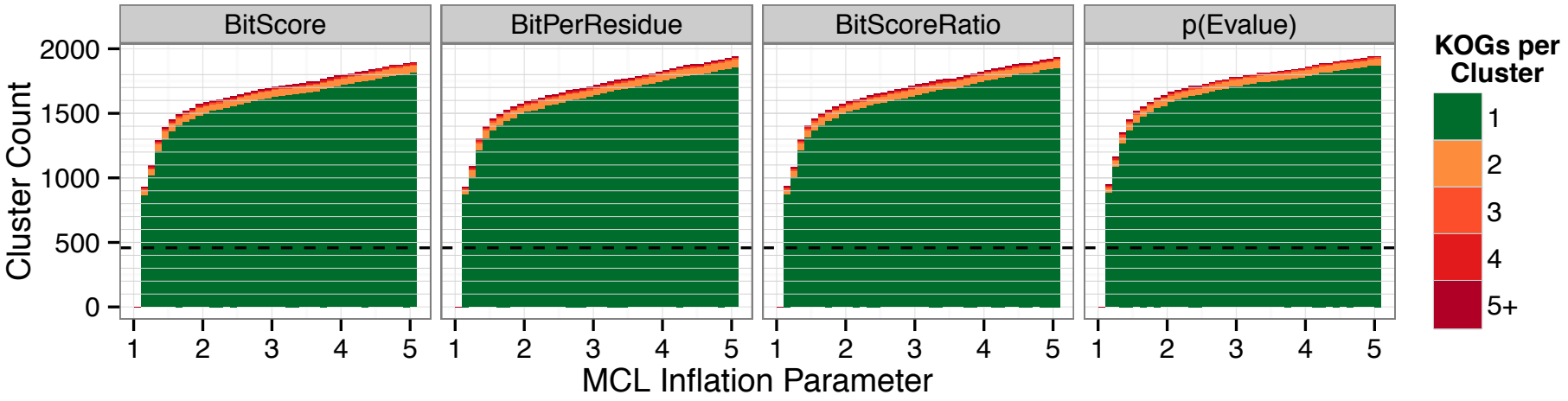
Sensitivity

eck_33333333333_shf_evn_1e−5_nrm_clusters_per_kog_summary.Rtab



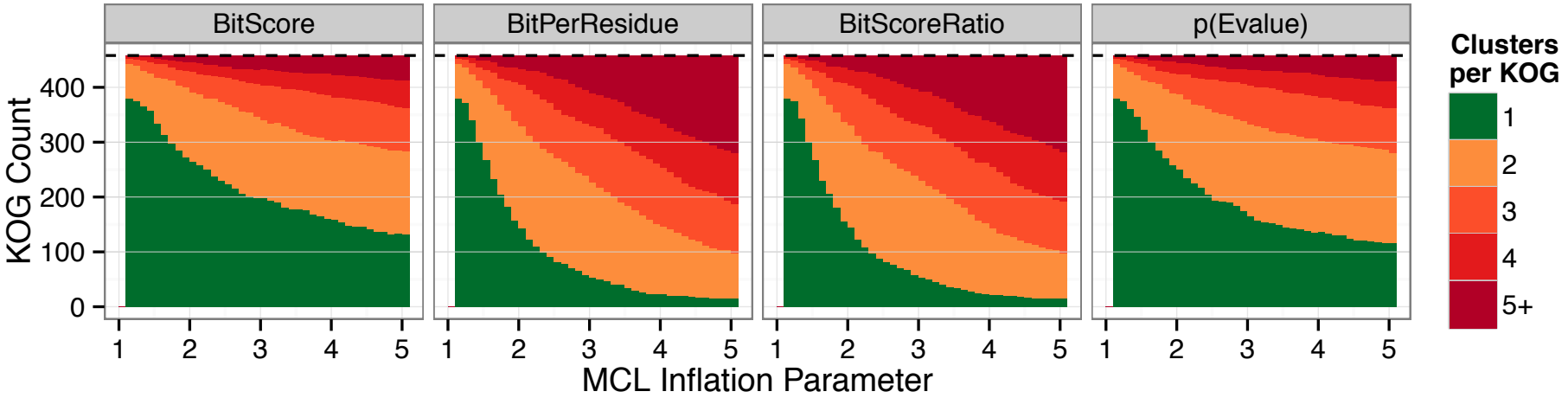
Specificity

eck_33333333333_shf_evn_1e−5_nrm_kogs_per_cluster_summary.Rtab



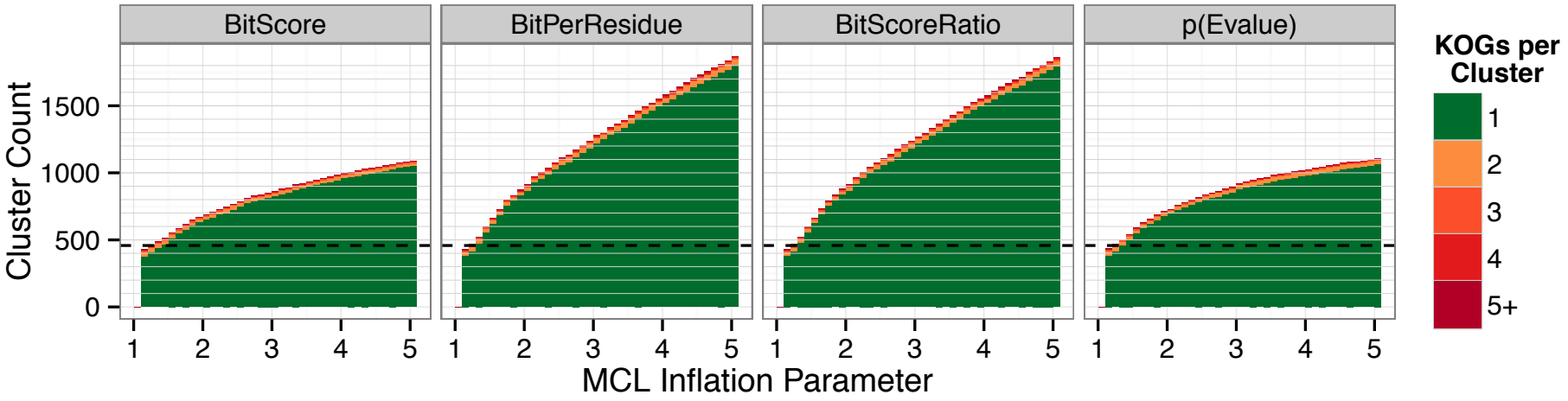
Sensitivity

eck_33333333333_shf_rnd_1e−5_nrm_clusters_per_kog_summary.Rtab



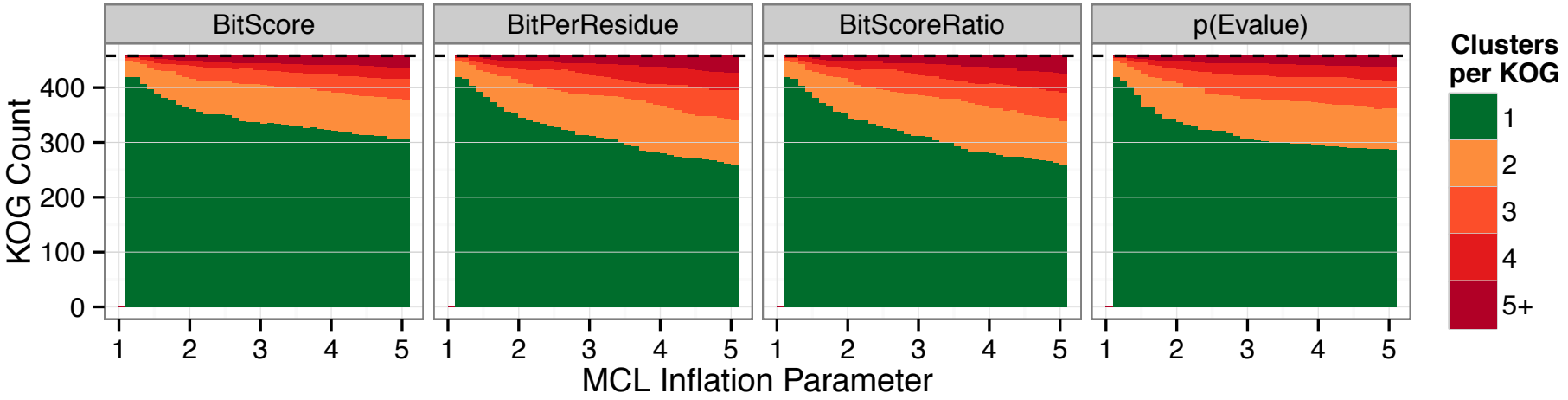
Specificity

eck_33333333333_shf_rnd_1e−5_nrm_kogs_per_cluster_summary.Rtab



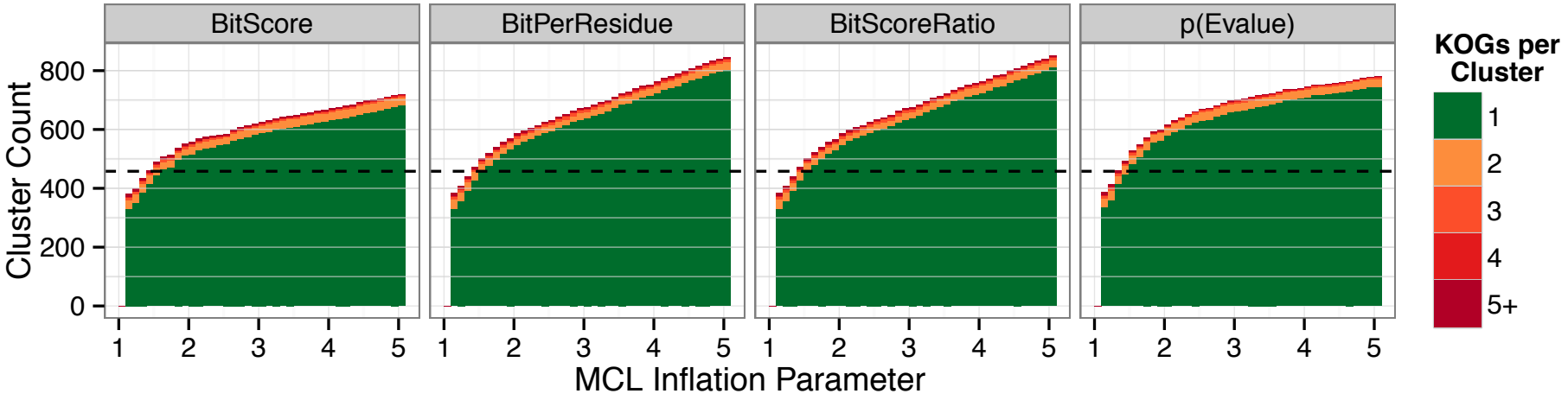
Sensitivity

eck_123/shf_evn/1e−5/nrm//eck_12312312312_shf_evn_1e−5_nrm_clusters_per_kog_summary.Rtab



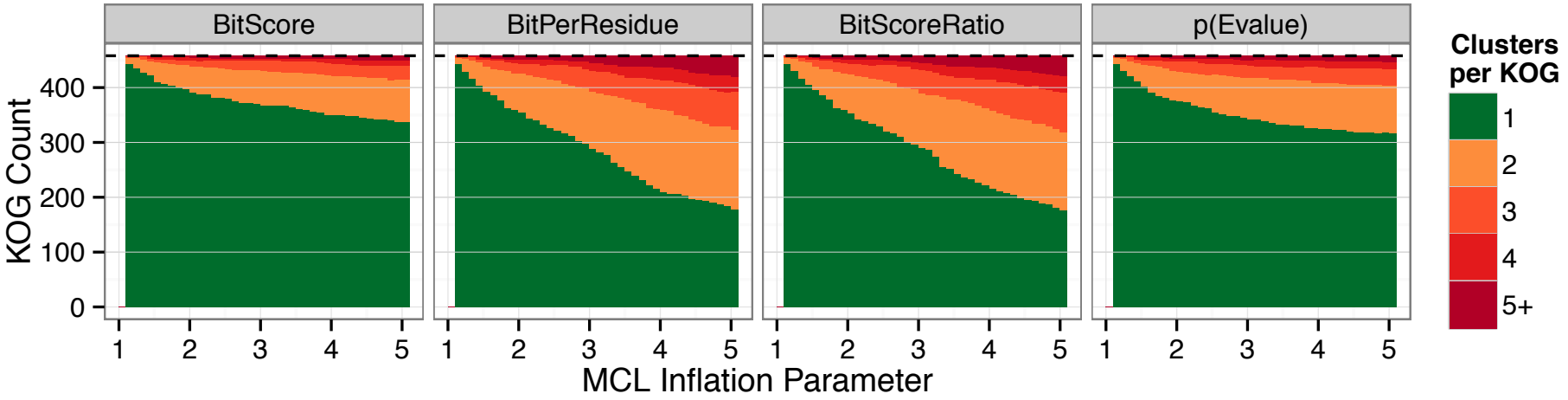
Specificity

eck_123/shf_evn/1e−5/nrm//eck_12312312312_shf_evn_1e−5_nrm_kogs_per_cluster_summary.Rtab



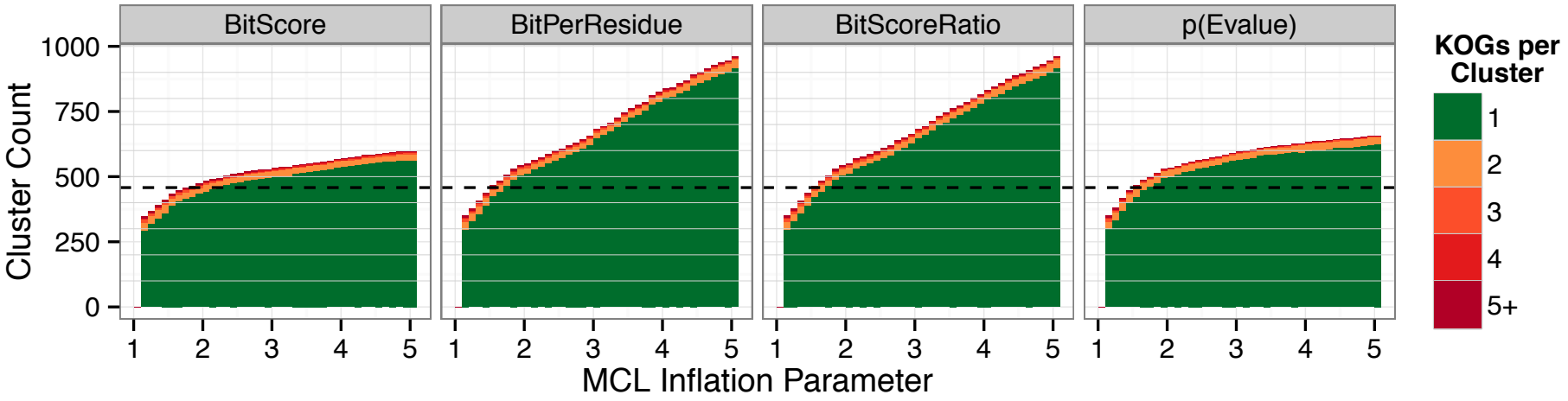
Sensitivity

eck_123/shf_rnd/1e−5/nrm//eck_12312312312_shf_rnd_1e−5_nrm_clusters_per_kog_summary.Rtab



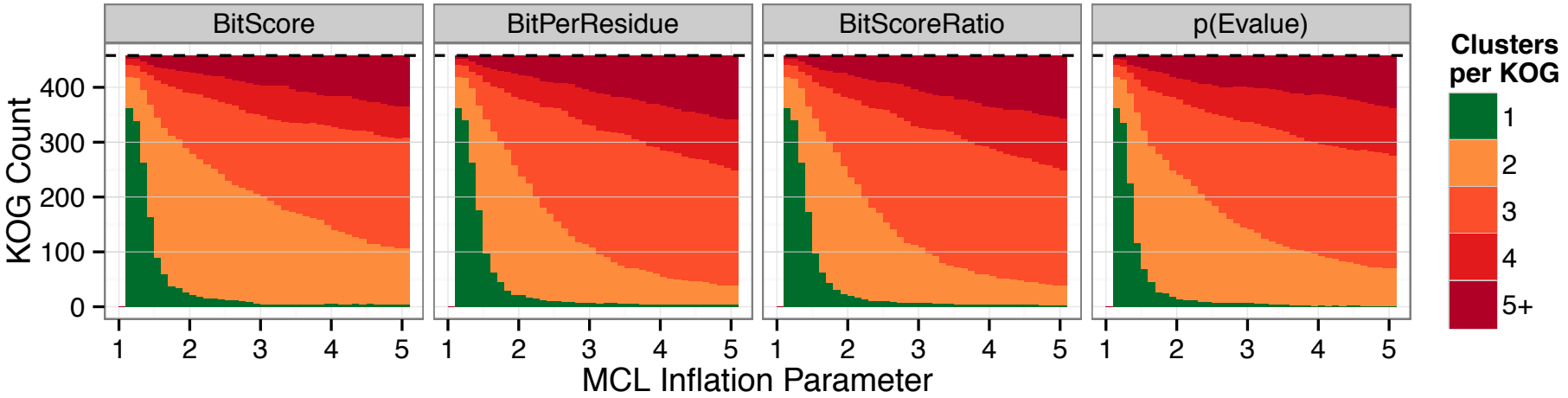
Specificity

eck_123/shf_rnd/1e−5/nrm//eck_12312312312_shf_rnd_1e−5_nrm_kogs_per_cluster_summary.Rtab



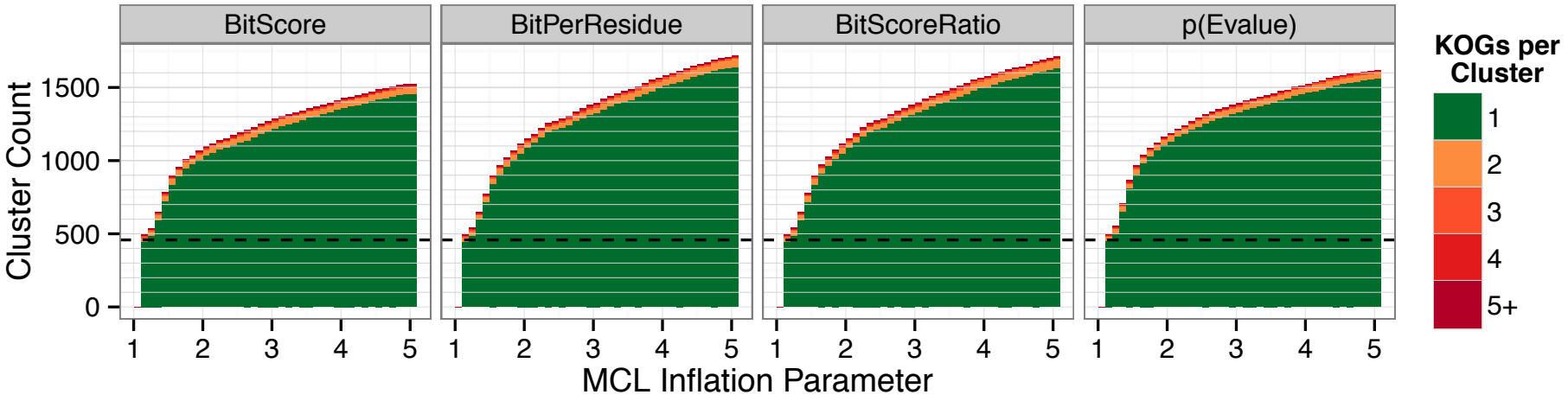
Sensitivity

eck_32332332332_shf_evn_1e−5_nrm_clusters_per_kog_summary.Rtab



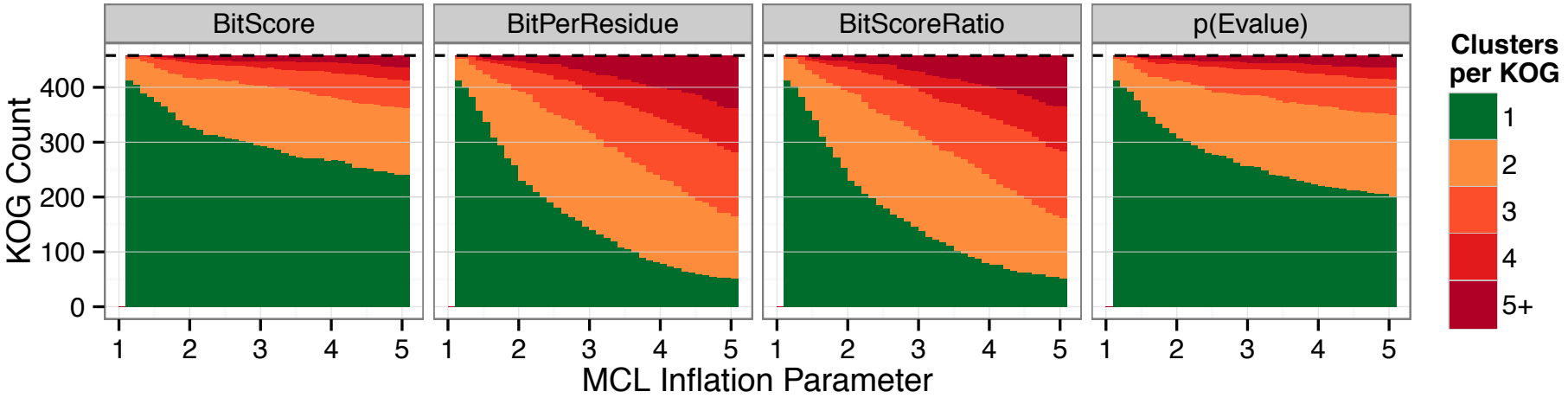
Specificity

eck_32332332332_shf_evn_1e−5_nrm_kogs_per_cluster_summary.Rtab



Sensitivity

eck_32332332332_shf_rnd_1e−5_nrm_clusters_per_kog_summary.Rtab



Specificity

eck_32332332332_shf_rnd_1e−5_nrm_kogs_per_cluster_summary.Rtab